Smart Grid using Big Data Analytics: A Random Matrix Theory Approach

Smart Grid using Big Data Analytics

A Random Matrix Theory Approach

Robert C. Qiu and Paul Antonik

WILEY

This edition first published 2017

Registered Offices
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Office
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloging-in-Publication Data

Names: Qiu, Robert C., 1966– author. | Antonik, Paul, author.
Title: Smart grid using big data analytics / Robert C. Qiu, Paul Antonik.
Description: Chichester, West Sussex, United Kingdom : John Wiley & Sons, Inc., 2017. |
Includes bibliographical references and index.
Identifiers: LCCN 2016042795| ISBN 9781118494059 (cloth) | ISBN 9781118716793 (epub) |
ISBN 9781118716809 (Adobe PDF)
Subjects: LCSH: Smart power grids. | Big data.
Classification: LCC TK3105 .Q25 2017 | DDC 621.310285/57–dc23
LC record available at https://lccn.loc.gov/2016042795

Cover design by Wiley
Cover image: johnason/loops7/ziggymaj/gettyimages

Set in 10/12pt Warnock by SPi Global, Pondicherry, India

10 9 8 7 6 5 4 3 2 1

To Lily L. Li

Contents

Preface

When, in the fall of 2010, the first author wrote the initial draft of this book in the form of lecture notes for a smart grid course, the preface began by justifying the need for such a course. He explained at length why it was important that electrical engineers understand Smart Grid. Now, such a justification seems unnecessary. Rather, he had to justify repeatedly his own decision to cover aspects of big data in a smart grid course, in order to convince the audience and most of the time himself. Although we feel completely comfortable with this "big" decision at this point of writing, we still want to outline some points that led to that decision. The decision was motivated by our passion to pursue research in this direction. The excitement of the problems that lie at the intersection between the two topics convinced us that the time had come to study big data for smart grid, which is the integration of communications and sensing.

For big data, we have two major tasks: (i) big data modeling and (ii) big data analytics. After the book was finished, we realized that more than 90% of the contents were dedicated to these two aspects. The applications of this material are treated very lightly. We emphasize the mathematical foundation of big data, in a similar way to Qiu and Wicks' *Cognitive Networked Sensing and Big Data* (Springer, 2014). Qiu, Hu, Li and Wicks' *Cognitive Radio Communication and Networking* (John Wiley & Sons Ltd, 2012) complements both books. All three books are unified by matrix-valued random variables (random matrix theory).

In choosing topics we heeded the warning of the former NYU professor K. O. Friedrichs: "It is easy to write a book if you are willing to put into it everything you know about the subject" (P. Lax, *Functional Analysis*, Wiley-Interscience, 2002, p. xvii). The services provided by Google Scholar and online digital libraries completely relieved us of the burden of physically going to the library. Using the "cited by" function provided by Google Scholar, even working remotely from the office, we could put things together without difficulty. We were able to use this function to track the latest results on the subject. This book deals with the fundamentals of big data, addressing principles and applications. We view big data as a new science: a combination of information science and data science. Smart grid, communications and sensing are three applications of special interest to the authors.

This book studies the intersection of big data (Part I) with Smart Grid (part II) and communications and sensing (Part III). Random matrix theory is treated as the unifying

theme. Random matrix models provide a powerful framework for modeling numerous physical phenomena in quantum systems, financial systems, sensor networks, wireless networks, smart grid, and so forth. One goal is to outline how an audience with a signals-and-systems background can contribute to big data research and development (R&D). As most mathematical results are synthesized from the literature of mathematics and physics, we have tried to present them in very different ways, usually motivated by the above Big Data systems. Roughly speaking, a big data system means a large *statistical* system or "large models." Although no claim of novel mathematical results is made, the combination of these mathematical models with these particular big data systems seems worth mentioning. Initially, we really intended to write a textbook in a traditional way; however, as the project evolved, we could not resist the temptation to include many beautiful mathematical results. These results are relatively new in the statistical literature and completely novel to the engineering community. We aim to bridge the gap between big data modeling/analytics and large random matrices in a systematic manner. The latest references are reasonably comprehensive in this treatment (sometimes exhaustive, for example for non-Hermitian random matrices).

Random matrices are ubiquitous [1]. The reason for this is twofold. First, they have a great degree of universality; that is, the eigenvalue properties of large matrices do not depend on the underlying statistical matrix ensemble. Second, random matrices can be viewed as noncommuting probability theory where the whole matrix is treated as an element of the probability space. Nowadays, data sets are usually organized as large matrices whose first dimension is equal to the number of degrees of freedom and the second to the number of measurements. Typical examples include financial systems, sensing systems and wireless communications systems.

As pointed out above, random matrix theory is the foundation for many problems in smart grid and big data. We hold the belief that big data is more basic than smart grid; the latter is is the applied science of the former. On the other hand, smart grid motivates big data. As a result, the close interaction is the natural topic for study. During the first offering of the first author's course on smart grid (Fall 2010), he primarily relied on the journal papers on power systems. During the second offering of this course (Fall 2013), the contents of the materials mainly covered big data aspects, especially the latest results of the random matrix theory. The audience were graduate students from EE and CS. He realized that without solid backgrounds in big data, the introduction of smart grid—large power systems that lead to high-dimensional data—could be very superficial. For example, the challenges of state estimation and bad data detection are due to the high dimensionality of the resultant datasets. This issue belongs to the larger class of standard big data problems. Although, in Fall 2013, he pushed the course to the frontiers of statistics, theoretical physics, and finance, he knew that his class had difficulty in following him. He lost most students when he addressed random matrices. To do that, he had to go back to cover random vectors first. It was a very painful experience for all of them because the students were not comfortable with random vectors, which are the most important prerequisite for reading Part I of this book.

Big data is a new science with numerous applications. After we combine smart grid and big data, we are able to crystallize many standard problems and focus our efforts on the marriage of two subjects. We feel very comfortable that this combination will be extremely fruitful in the near future. In the infancy of this connection, our aim is to

spell out our goals and methodologies; at the same time, we outline the mathematical foundations by introducing random matrix theory, in the hope that this mathematical theory is sufficiently general and flexible to provide a definitive machinery for the analysis of big data and smart grid. It is common to hear that big data lacks a theoretical foundation. Maybe there is no theory at all. It is the sense of the mission (to search for such a theory) that has sustained us in this long journey.

Acknowledgments

This book is the result of many years of teaching and research in the field of smart grid. This work is in part funded by the National Science Foundation through three grants (ECCS-0901420, ECCS-0821658, and CNS-1247778), and the Office of Naval Research through two grants (N00010-10-1-0810 and N00014-11-1-0006). We want to thank Dr. Santanu K. Das (ONR) for his support for the work.

We want to thank Dr. Zhen Hu for reading through the whole manuscript. The first author wants to thank the ECE students in his smart grid courses (Fall 2010, Fall 2013) for their patience and useful feedback. The first author was working with China Power Research Institute (CPRI), Beijing, China, when the book was nearing completion. He wants to thank his host Dr. Dongxia Zhang (CPRI) and Dr. Chaoyang Zhu for their hospitality. The first author also worked for two months at Shanghai Jiaotong University. He wants to thank Professors Wenxian Yu, Xiuchen Jiang and Zhijian Jin for their hospitality and useful discussions. He also wants to thank Professors Shaoqian Li and Guangrong Yue at the University of Electronic Science and Technology of China (UESTC) for their hospitality and useful discussions.

Some Notation

$\mathbf{A} \otimes \mathbf{B}$	Kronecker product of two matrices
$\mathbf{A}\,(p \times q)$	matrix with p rows and q columns
$\boxplus$	free additive convolution; Voiculescu's operations $\boxplus$ and $\boxtimes$
$\boxtimes$	free multiplicative convolution
$\langle \cdot \rangle$	expectation of $\cdot$
$\|\cdot\|_{\mathrm{op}}$	operator norm of the matrix
$\|\cdot\|_F$	Frobenius norm of the matrix
$\xrightarrow{D}$	convergence in distribution
$\mathbb{C}$	the set of complex numbers
$\mathbb{C}^+$	$\mathbb{C}^+ = \{z \in \mathbb{C} : \mathrm{Im}\,(z) > 0\}$
$\mathbb{E}X$	expectation of random variable X
$\mathbb{E}x$	expectation of random vector x
$\mathbb{E}X$	expectation of random matrix X
$\mathbb{I}_{x \in A}$	indicator function $\mathbb{I}_{x \in A}$ is 1 if the event $x \in A$ is true
Im	imaginary part of real number z
$m(z)$	Stieltjes transform
$\mathbb{N}$	set of natural numbers
$\mathbb{P}$	probability
$\mathbb{R}$	set of real numbers
$\mathbb{R}^+$	set of positive real numbers
$\mathrm{Re}\,(z)$	real part of real number z
$\mathbb{Z}$	set of integer numbers

1

Introduction

1.1 Big Data: Basic Concepts

Data is "unreasonably effective" [2]. Nobel laureate Eugene Wigner referred to the unreasonable effectiveness of mathematics in the natural sciences [3]. What is big data? According to [4], its sizes are in the order of terabytes or petabytes; it is often online, and it is not available from a central source. It is diverse, may be loosely structured with a large percentage of data missing . It is heterogeneous.

The promise of data-driven decision-making is now broadly recognized [5–16]. There is no clear consensus about what big data is. In fact, there have been many controversial statements about big data, such as "Size is the only thing that matters."

Big data is a big deal [17]. The Big Data Research and Development Initiative has been launched by the US Federal government. "By improving our ability to extract knowledge and insights from large and complex collections of digital data, the initiative promises to help accelerate the pace of discovery in science and engineering, strengthen our national security, and transform teaching and learning" [17]. Universities are beginning to create new courses to prepare the next generation of "data scientists."

The age of big data has already arrived with global data doubling every two years. The utility industry is not the only one facing this issue (Wal-Mart has a million customer transactions a day) but utilities have been slower to respond to the data deluge. Scaling up the algorithms to massive datasets is a big challenge.

According to [18]:

> A key tenet of big data is that the world and the data that describe it are constantly changing and organizations that can recognize the changes and react quickly and intelligently will have the upper hand ... As the volume of data explodes, organizations will need analytic tools that are reliable, robust and capable of being automated. At the same time, the analytics, algorithms, and user interfaces they employ will need to facilitate interactions with the people who work with the tools.

1.1.1 Big Data—Big Picture

Data is a strategic resource, together with natural resources and human resources. Data is king! "Big data" refers to a technology phenomenon that has arisen since the late 1980s [19]. As computers have improved, their growing storage and processing capacities have provided new and powerful ways to gain insight into the world by sifting

Smart Grid using Big Data Analytics: A Random Matrix Theory Approach, First Edition.
Robert C. Qiu and Paul Antonik.

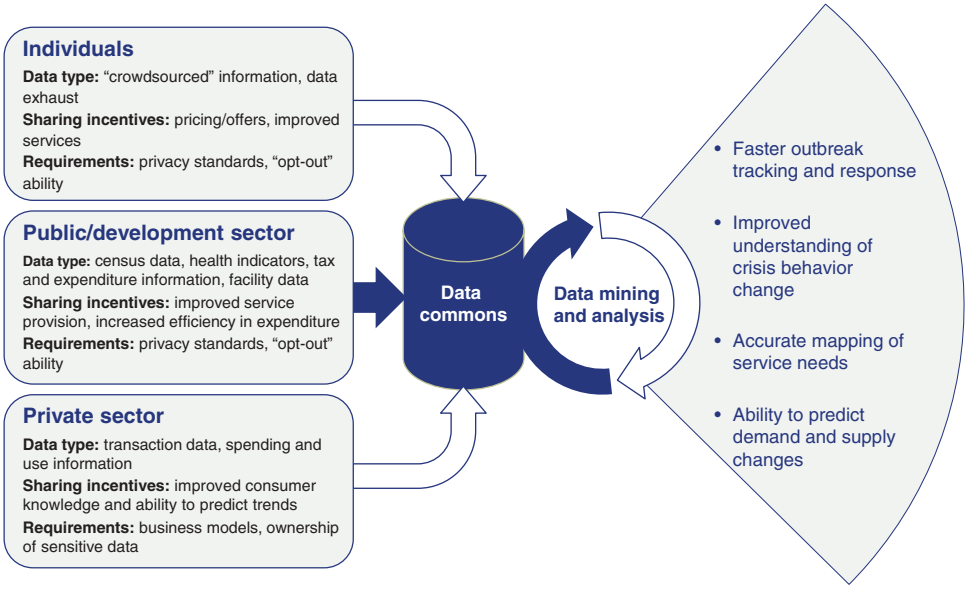


Figure 1.1 Big data, big impact: new possibilities for international development. Source: Reproduced from [6] with permission from the World Economic Forum.

through enormous quantities of data available. But this insight, discoverable in previously unseen patterns and trends within these phenomenally large data sets, can be hard to detect without new analytic tools that can comb through the information and highlight points of interest.

Sources such as online or mobile financial transactions, social media traffic, and GPS coordinates, now generate over 2.5 quintillion bytes of so-called "big data" every day. The growth of mobile data traffic from subscribers in emerging markets exceeded 100% annually through 2015. There are new possibilities for international development (see Figure 1.1).

Big data at the societal level provides a powerful microscope, together with social mining—the ability to discover knowledge from these data. Scientific research is being revolutionized by this, and policy making is next in line, because big data and social mining are providing novel means for measuring and monitoring wellbeing in our society more realistically, beyond the GDP, more precisely, continuously, everywhere [20].

Most scientific disciplines are finding the data deluge to be extremely challenging, and tremendous opportunities can be realized if we can better organize and access the data [16].

Chris Anderson believed that the data deluge makes the scientific method obsolete [21]. Petabytes data tell us to say correlation is enough. There is no need to find the models. Correction replaces causality. It remains open to see whether the data growth will lead to a fundamental change in scientific methods.

In the computing industry we are now focussing on how to process big data [22].

A fundamental question is "What is the unifying theory for big data?" This book adopts the viewpoint that big data is a new science of combining data science and information

science. Specialists in different fields deal with big data on their own, while information experts play a secondary role as assistants. In other words, most scientific problems are in the hands of specialists whereas only few problems—common to all fields—are refined by computing experts. When more and more problems are open, some unifying challenges common to all fields will arise. Big data from the Internet may receive more attention first. Big data from physical systems will become more and more important.

Big data will form a unique discipline that requires expertise from mathematics, statistics and computing algorithms.

Following the excellent review in [22], we highlight some challenges for big data:

- *Processing unstructured and semistructured data.* Presently 85% of the data are unstructured or semistructured. Traditional relational databases cannot handle these massive datasets. High scalability is the most important requirement for big-data analysis. MapReduce and Hadoop are two nonrelational data analysis technologies.
- *Novel approaches for data representation.* Current data representation cannot visually express the true essence of the data. If the raw data are labeled, the problem is much easier but customers do not approve of the labeling.
- *Data fusion.* The true value of big data cannot exhibit itself without data fusion. The data deluge on the Internet has something to do with data formats. One critical challenge is whether we can conveniently fuse the data from individuals, industry and government. It is preferable that data formats be platform free.
- *Redundancy reduction and high-efficiency, low-cost data storage.* Redundancy reduction is important for cost reduction.
- *Analytical tools and development environments that are suitable for a variety of fields.* Computing algorithm researchers and people from different disciplines are encouraged to work together closely as a team. There are enormous barriers for people from different disciplines to share data. Data collection, especially simultaneous collection for relational data, is still very challenging.
- *Novel approaches to save energy for data processing, data storage, and communication.*

1.1.2 DARPA's XDATA Program

The Defense Advanced Research Projects Agency's (DARPA's) XDATA program seeks to develop computational techniques and software tools for analyzing large volumes of data, both semistructured (e.g., tabular, relational, categorical, metadata) and unstructured (e.g., text documents, message traffic). Central challenges to be addressed include (i) developing scalable algorithms for processing imperfect data in distributed data stores, and (ii) creating effective human–computer interaction tools to facilitate rapidly customizable visual reasoning for diverse missions.

Data continues to be generated and digitally archived at increasing rates, resulting in vast databases available for search and analysis. Access to these databases has generated new insights through data-driven methods in the commercial, science, and computing sectors [23]. The defense section is "swimming in sensors and drowning in data." Big data arises from the Internet and the monitoring of industrial equipment. Sensor networks and the Internet of Things (IoT) are another two drivers.

There is a trend for data to be used that can sometimes be seen only once, for milliseconds, or can only be stored for a short time before being deleted, especially in some defense applications. This trend is accelerated by the proliferation of various digital devices and the Internet. It is important to develop fast, scalable, and efficient methods for processing and visualizing data.

The XDATA program's technology development is approached through four technical areas (TAs):

- TA1: Scalable analytics and data-processing technology;
- TA2: Visual user interface technology;
- TA3: Research software integration;
- TA4: Evaluation.

It is useful to consider distributed computing via architectures like MapReduce, and its open source implementation, Hadoop. Data collected by the Department of Defense (DoD) are particularly difficult to deal with, including missing data, missing connections between data, incomplete data, corrupted data, data of variable size and type, and so forth [23]. We need to develop analytical principles and implementations *scalable* to data volume and distributed computer architectures. The challenge for Technical Area 1 is how to enable systematic use of big data in the following list of topic areas:

- Methods for leveraging the problem structure to create new algorithms to achieve optimal tradeoffs among time complexity, space complexity, and stream complexity (i.e., how many passes over the data are needed).
- Methods for the propagation of uncertainty (i.e., every query should have an answer and an error bar), with performance guarantees for loss of precision due to approximations.
- Methods for measuring nonlinear relationships among data.
- Sampling and estimation techniques for distributed platforms, including compensating for missing information, corrupted information, and incomplete information.
- Methods for distributed dimensionality reduction, matrix factorization, matrix completion (within a distributed data store where data are not all in one place).
- Methods for operating on streaming data feeds.
- Methods for determining optimal cloud configurations and resource allocation with asymmetric components (e.g., many standard machines, a small number of large-memory machines, machines with graphical processing units).

The challenge for Technical Area 2 is how to hook up big data analytics to interfaces, including but not limited to the following topics:

- Visualization of data for scientific discovery, activity patterns, and summaries.
- Expressive visualization and/or query languages and processing that support domain-specific interaction, successive query refinement, repeated viewing of data, faceted search, multidimensional queries, and collaborative/interactive search.
- Principled design, including menus, query boxes, hover tips, invalid action notifications, layout logic, as well as processes of overview, zoom and filter, and details-on-demand.
- Support for the study and characterization of users, including extraction of relations and history, usage, hover time, click rate, dwell, etc.

- Functions of timeliness, online versus batch processing, metainformation, etc.
- Analytical workflows including data cleaning and intermediate processing.
- Tools for rapid domain-specific end-user customization.

1.1.3 National Science Foundation

The phrase "big data" in the National Science Foundation (NSF) refers to large, diverse, complex, longitudinal, and/or distributed data sets generated from instruments, sensors, Internet transactions, e-mail, video, click streams, and/or all other digital sources available today and in the future [5].

Today, US government agencies recognize that the scientific, biomedical and engineering research communities are undergoing a profound transformation with the use of large-scale, diverse, and high-resolution data sets that allow for data-intensive decision making, including clinical decision making, at a level never before imagined. New statistical and mathematical algorithms, prediction techniques, and modeling methods, as well as multidisciplinary approaches to data collection, data analysis and new technologies for sharing data and information are enabling a paradigm shift in scientific and biomedical investigation. Advances in machine learning, data mining, and visualization are enabling new ways of extracting useful information in a timely fashion from massive data sets, which complement and extend existing methods of hypothesis testing and statistical inference. As a result, a number of agencies are developing big-data strategies to align with their missions. The NSF's solicitation focuses on common interests in big data research across the National Institutes of Health (NIH) and the NSF.

1.1.4 Challenges and Opportunities with Big Data

There are challenges with Big Data. The first step is data acquisition. Some data sources, such as sensor networks, can produce staggering amounts of raw data. A lot of this data is not of interest. It can be filtered out and compressed by orders of magnitude. One challenge is to define these filters in such a way that they do not discard useful information.

The second big challenge is to generate the right metadata automatically, and to describe what data is recorded and how it is recorded and measured. This metadata is likely to be crucial to downstream analysis. Frequently, the information collected will not be in a format ready for analysis. We have to deal with erroneous data: some news reports are inaccurate.

Data analysis is considerably more challenging than simply locating, identifying, understanding, and citing data. For effective large-scale analysis all of this has to happen in a completely automated manner.

Mining requires integrated, cleaned, trustworthy, and efficiently accessible data, declarative query and mining interfaces, scalable mining algorithms, and big data computing environments. Today's analysts are impeded by a tedious process of exporting data from the database, performing a non-SQL process, and bringing the data back.

Having the ability to analyze big data is of limited value if users cannot understand the analysis. Ultimately, a decision maker, provided with the result of analysis, has to interpret these results.

In short, there is a multistep pipeline required to extract value from data. This pipeline is not a simple linear flow—rather, there are frequent loops back as downstream steps suggest changes to upstream steps.

There has not been a commonly accepted definition of big data. In [24], there are some claims that may define the ballpark:

- Big data is the same as scalable analytics.
- Big-data problems are primarily on the application side.
- Big-data problems are primarily at the systems level.
- Big-data requires a cloud-based platform.
- The data-management community is in danger of missing the big-data train.
- It is not possible to conduct big-data research effectively without collaborating with people outside the data-management community.
- All the big-data problems can be reduced to MapReduce problems [25].
- The bulk of big-data challenges are being addressed by industry.
- The bulk of big-data challenges are at the implementation level.
- Size is the only thing that matters (for big data).

The growth of the data volume seems to outspend the advance of our computing infrastructure. Conventional data-processing technologies, such as database and data warehouse, are becoming inadequate for the amount of data.

1.1.5 Signal Processing and Systems Engineering for Big Data

The big-data workshop for signal processing and systems engineering was held in 2013 [4]. One motivation from the NSF's point of view [4] is to leverage analytical, computational, storage, and implementation tools:

- assess fundamental performance limits in processing and storage;
- develop scalable algorithms: online (adaptive) and decentralized;
- complement computer and information science and engineering (CISE) efforts on parallel architectures and computing;
- account for redundancy and error control: source and channel (de)coding;
- cross-fertilize NSF-wide advances on fault-tolerance, privacy, and security.

Another motivation is to facilitate ground-breaking research in big-data science and engineering:

- to offer top-down approaches for signal processing and systems engineering;
- to develop a toolbox for statistics and optimization.

High-level issues of interest include: Can lessons learned from "big systems" engineering be applied to big-data engineering? What are the right pathways? What are overarching tools to catalyze big-data collaboration between scientists and engineers? What are the grand challenges in big data science and engineering? How should we educate engineers about big data?

Big *engineering* data has unique characteristics: it is more disciplined and regulated. There are emerging engineering systems with big-data opportunities: smart grids, sensor nets, transportation, telemedicine, aerospace, testing, safety, nuclear, design blueprints and more. Now it may be necessary to rethink data collection and storage to facilitate big-data processing and inference tasks.

Some sample questions are: How do we trade off complexity for accuracy in massive decentralized signal and data analysis tasks? How can efficient signal and data analysis

algorithms be developed for big, unstructured or loosely structured data? What are the basic principles and useful methodologies to scale inference and learning algorithms and trade off the computational resources (e.g., time, space and energy) according to the needs of engineering practice (e.g. robustness versus efficiency, real time)?

Big-data processing and analysis, according to Hero [26], require the following: (i) Integration of very heterogenous data: correlation mining in massive database; processing data at vastly different scales and noise levels; processing a mixture of continuous and categorical variables. (ii) Reliable and robust quantitative models: uncertainty quantification; adaptation to drift over time. (iii) High throughput real-time processing: smart adaptive sampling and compression; distributed or parallel processing architectures. (iv) Interactive user interfaces: human-in-the-loop processing; visualization and dimensionality reduction.

Some signal-processing challenges, according to Hero [26], include the following. (i) Heterogeneous data integration: ranking signals for human-aided selection of relevant variables; fusing graphs, tensors, and sequence data; active visualization: dimensionality reduction. (ii) Flexible low-complexity modeling and computation: scalable signal processing: distributed algorithms and implementation; smart sampling: feedback-controlled signal search and acquisition. (iii) Reliable robust models for anomaly detection and classification: parsimonious signal processing; sparse correlation graphical models; decomposable signal processing: factored models and algorithms.

As for the signal-processing toolbox, we have the following primitives: linear equation solvers (Gauss, Givens, Householder); spectral representations (FFT, SVD); ensemble averaging (cross validation, bootstrap, boosting); optimization (linear least square, linear and quadratic programming, dynamic programing). They can used for the following applications. (i) Linear and nonlinear prediction: Wiener, Kalman, particle filtering, Volterra filters; (ii) signal reconstruction: matrix factorization, matrix, completion, robust principal component analysis (PCA). (iii) Dimension reduction: PCA, independent component analysis (ICA), independent principal component analysis (IPCA), canonical correlation analysis (CCA), linear discriminant analysis (LDA), nonlinear editing (NLE). (iv) Adaptive sampling: compressive sensing, distilled sensing, sketching. (v) Signal processing on graphs: graph spectra, the k-nearest-neighbor algorithm (k-NN) search, belief propagation.

There is a growing gap between the amount of data we generate and the amount of data we are able to store, communicate, and process. As Richard Baraniuk points out, we have produced already twice as much data as can be stored [27]. And the gap keeps widening. As long as this continues there is an urgent need for novel data-acquisition concepts like compressive sensing.

Compressive sensing and sparse representations play a key role: advanced probability theory and (in particular) random matrix theory, convex optimization, and applied harmonic analysis are becoming standard ingredients of the toolbox of many engineers. Compressive sensing has advanced the development of ℓ_1-minimization algorithms, and more generally of nonsmooth optimization. These algorithms find widespread use in many disciplines, including physics, biology, and economics [28]. The most important legacy of compressive sensing may be that it has forced us to think about information, complexity, hardware, and algorithms in a truly integrated manner.

Nondominated sorting is an interesting and useful framework for multicriteria anomalies, human-machine interaction, or multiple end users [29, 30].

The author's research proposals to the National Science Foundation (NSF) [31–34] are relevant in the context of this section.

1.1.6 Large Random Matrices for Big Data

Random matrices play a central role in statistics in the context of multivariate data. Three classical books are included here [35–37].

The continued growth of big data has given rise to high-dimensional statistical analysis. Convex analysis, Riemannian geometry and combinatorics are relevant. Random matrix theory (RMT) has emerged as a particularly useful framework for many theoretical questions associated with the analysis of high-dimensional multivariate data; see [38] for a recent overview of RMT.

RMT affects modern statistical thinking in two ways. On one hand, most of the mathematical treatments of RMT have focused on matrices with a high degree of independence in the entries, which one may refer to as "unstructured" random matrices. Recall that about 75% of big data is unstructured. On the other hand, in high-dimensional statistics, we are primarily interested in problems where there are lower dimensional structures buried under random noise.

In November 2011, the author of [39] dedicated over 200 pages to random matrix theory. In [40], the whole book was motivated by the same vision but with different regimes. The first book deals with so-called asymptotic regimes, while the second deals with nonasymptotic regimes. In the asymptotic regimes, the sizes of random matrices are assumed to approach infinity. For example, for a random matrix of $\mathbf{X}$ of size $m \times n$, we assume the asymptotic regime: $m \to \infty$, $n \to \infty$, but $m/n \to c$. On the other hand, the nonasymptotic regime is defined as: m and n are large, but *finite*. The author's research proposals to the National Science Foundation (NSF) [31–34] have a similar motivation.

As pointed out in Section 1.1.1, "High scalability is the most important requirement for big data analysis." Some state "Size is the only thing that matters." Based on this observation, it seems natural to the author to model big data using a nonasymptotic theory of random matrices. The motivation is to investigate how the algorithms *scale* with sizes of data samples.

We believe that a nonasymptotic theory of random matrices can unify many big-data problems. It is our intention to use this theory as the departure point for many problems studied later in this book.

Tensors (also known as multidimensional arrays or N-way arrays) are used in a variety of applications ranging from chemometrics to network analysis. The Tensor Toolbox [41] provides classes for manipulating dense, sparse, and structured tensors using MATLAB's object-oriented features.

1.1.7 Big Data Across the US Federal Government

We highlight some points [42] that which are relevant to the contex of this book.

The **Anomaly Detection at Multiple Scales** program at DARPA creates, adapts and applies technology to anomaly characterization and detection in massive data sets. Anomalies in data cue the collection of additional, actionable information in a wide variety of real-world contexts. The initial application domain is insider threat

detection in which malevolent (or possibly inadvertent) actions by a trusted individual are detected against a background of everyday network activity.

The Department of Energy (DOE) provides leadership to the data management, visualization and data analytics communities, including digital preservation and community access. **Mathematics for Analysis of Petascale Data** addresses the mathematical challenges of extracting insight from huge scientific datasets, finding key features and understanding the relationships between those features. Research areas include machine learning, real-time analysis of streaming data, stochastic nonlinear data-reduction techniques, and scalable statistical analysis techniques applicable to a broad range of DOE applications including sensor data from the electric grid, cosmology, and climate data.

The **Office of Basic Energy Sciences (BES) BES Scientific User Facilities** have supported a number of efforts aimed at assisting users with data management and analysis of big data, which can be as big as *terabytes* (1012 bytes) of data *per day* from a single experiment.

Researchers funded by the NSF are developing a unified theoretical framework for principled statistical approaches to network models with scalable algorithms in order to differentiate knowledge in a network from randomness.

Information Integration and Informatics funded by the NSF addresses the challenges and scalability problems involved in moving from traditional scientific research data to very large, heterogeneous data, such as the integration of new data types models and representations, as well as issues related to data path, information life-cycle management, and new platforms.

NSF funds a distinct discipline encompassing mathematical and statistical foundations and computational algorithms. High-speed networks distribute over 15 petabytes of data each year in real time from the Large Hadron Collider (LHC) at CERN in Switzerland to more than 100 computing facilities.

The **Theoretical and Computational Astrophysics Networks (TCAN)** program seeks to maximize the discovery potential of massive astronomical data sets by advancing the fundamental theoretical and computational approaches needed to interpret those data, uniting researchers in collaborative networks that cross institutional and geographical divides, and training the future theoretical and computational scientists.

There are research projects (i) developing data visualizations in the defense of massive computer networks, and (ii) transforming big data sets and big ideas about earth science theories into scientific discoveries.

1.2 Data Mining with Big Data

Big data concern large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and data-collection capacity, big data is now rapidly expanding in all science and engineering domains, including physical, biological, and biomedical sciences.

Data collection has grown tremendously and is beyond the ability of commonly used software tools to capture, manage, and process within a "tolerable elapsed time." The most fundamental challenge for big data applications is to explore the large volumes of

data and extract useful information or knowledge for future actions [43]. In many situations, the knowledge-extraction process has to be very efficient and close to real time because storing all observed data is nearly infeasible. For example, the in-network processing in a large-scale cognitive radio network [44] is the bottleneck. For one microsecond of data collection, the processing time is in the level of several milliseconds (three orders of magnitudes larger). As a result, the unprecedented data volumes require an effective data analysis and prediction platform to achieve fast response and real-time classification for such big data.

Theorem 1.2.1 (HACE theorem [45]) Big data starts with large-volume, **h**eterogeneous, **a**utonomous sources with distributed and decentralized control, and seeks to explore **c**omplex and **e**volving relationships among data.

In the analogy of the blind men and the giant elephant, the localized (limited) view of each blind man leads to a biased conclusion. Exploring big data is equivalent to aggregating heterogeneous information from different sources (the blind men) to help draw a best possible picture to reveal the elephant in real time.

One of the fundamental characteristics of big data is the huge volume of data represented by heterogeneous and diverse dimensionalities. The reason is that different information collectors prefer their own schemata or protocols for data recording, and the nature of different applications also results in diverse data representations.

Autonomous data sources with distributed and decentralized controls are an important characteristic of big data applications. Being autonomous, each data source (a sensor) is able to generate and collect information without involving (or relying on) any centralized control.

While the volume of the big data increases, so do the complexity and the relationships underneath the data. One example is the time-varying wireless network or electric power grid.

A big-data processing framework is shown in Figure 1.2. The challenges at Tier I focus on data accessing and arithmetic computing procedures. Because big data is often stored at different locations and data volumes may continuously grow, an effective computing platform will have to take distributed large-scale data storage into consideration for computing. For example, distributed detection and estimation [46] is relevant in the context of wireless sensor networks.

Example 1.2.2 (a long time series) We form a large random matrix using a long record of time series. Given a time series $x[i]$, $i = 1, \ldots, NT$, where N and T are integers, we form a large random matrix $\mathbf{X}$ of $N \times T$. For example, $N = 1000$ and $T = 4000$. We view the data as a number of data segments. Here we have N data segments; the length of each segment is T, so a total of NT data samples are needed. □

Example 1.2.3 (the square kilometer array (SKA)—a big-data viewpoint) The square kilometer array (SKA) (see Figure 1.3) has 2000–3000 dishes. The wavelength ranges from 3 m to 3 cm. The SKA will have an array of coherently connected antennas spread over an area about 3000 km in extent, with an aggregate antenna collecting area of up to 106 m^2 at centimeter and meter wavelengths. The project timeline has the telescope operational below 10 GHz by 2022.

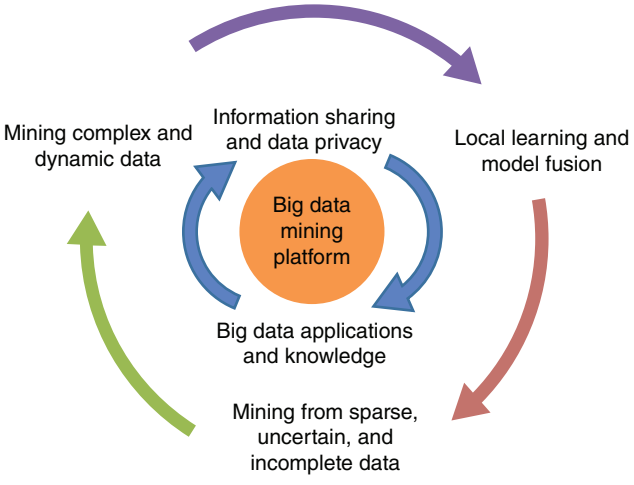


Figure 1.2 A big data processing framework. The research challenges form a three-tier structure and center around the "big data mining platform" (Tier I), which focuses on low-level data accessing and computing. Challenges on information sharing and privacy, and Big Data application domains and knowledge form Tier II, which concentrates on high-level semantics, application-domain knowledge, and user privacy issues. The outmost circle shows Tier III challenges on actual mining algorithms. Source: Reproduced from [45] with permission.

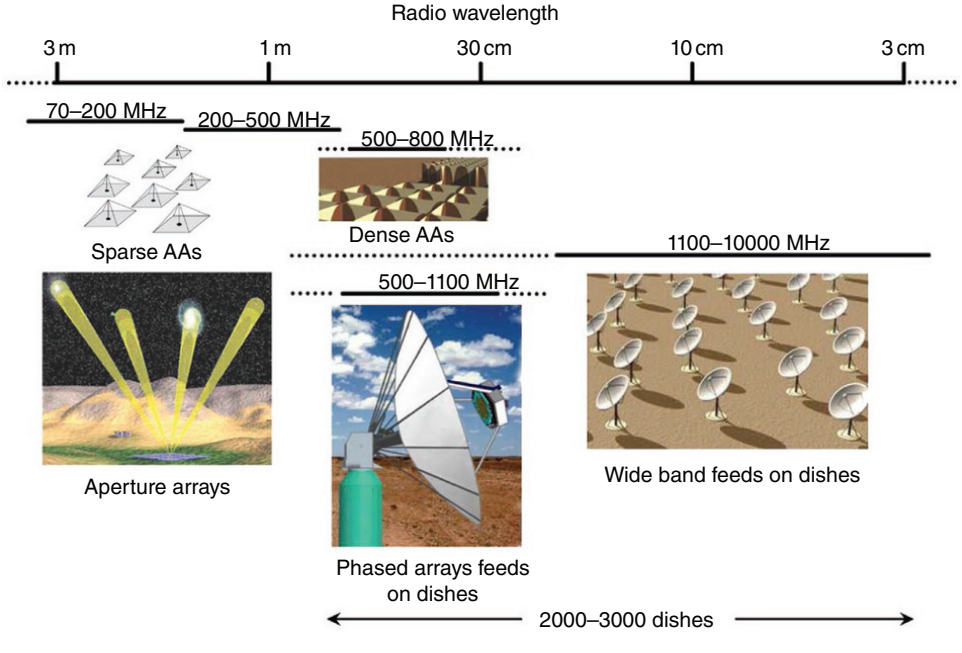


Figure 1.3 The square kilometer array. Source: Reproduced with permission from [47].

A large-scale wireless communication network is attempted to emulate a virtual array. So the SKA provides guidance.

With a 40 GB/s data volume, the data generated from the SKA are exceptionally large. We can model each dish as a sensor. So we deal with $N = 2000 - 3000$ sensors, which are spatially distributed. For each sensor, we observe a time series $\mathbf{x}_i \in \mathbb{C}^{T\times 1}$, for $i = 1, 2, \ldots, N$. We can collect the data from the N sensors into one single large matrix:

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_N^T \end{pmatrix}_{N\times T} \in \mathbb{C}^{N\times T}$$

The data for the SKA with time of T (called a snapshot) is represented by a large random matrix $\mathbf{X} \in \mathbb{C}^{N\times T}$. Now we study the time evolution of the data in a sequence of random matrices (for n snapshots) $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{C}^{N\times T}$. We can do some data processing using these large random matrices. (i) the sum of Hermitian random matrices (See Theorem 17.4.1) $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\mathbf{X}_1 \mathbf{X}_1^H + \ldots + \mathbf{X}_n \mathbf{X}_n^H \right)$ (ii) the product of non-Hermitian random matrices $\mathbf{X}_1 \cdots \mathbf{X}_n$; (iii) the geometric mean $\left(\mathbf{X}_1 \cdots \mathbf{X}_n \right)^{1/n}$. (iv) For N spatially distributed sensors (randomly), we form the data matrix $\mathbf{X}$ as above. What is the theoretical distribution of $\mathbf{X}$? It appears that this problem can be formulated in terms of a Euclidean random matrix. This problem corresponds to a random Green's function.

The so-called Euclidean random matrices, defined in Section 6.14, are a special class of random matrices. See also Section 16.1.5 for its connection with random geometric graphs. The elements A_{ij} of an $N \times N$ Euclidean random matrix $\mathbf{A}$ are given by a *deterministic* function f of positions of pairs of points that are randomly distributed in a finite region V of Euclidean space:

$$A_{ij} = f\left(\mathbf{r}_i, \mathbf{r}_j\right), \quad i, j = 1, \ldots, N$$

Here, the N points $\mathbf{r}_i$ are randomly distributed inside some region V of the d- dimensional Euclidean space with a uniform density $\rho = N/V$. □

Example 1.2.4 (local learning and model fusion for multiple information sources) As big data applications are featured with autonomous sources and decentralized controls, *aggregating distributed data sources* to a centralized site for mining is systematically prohibitive due to the potential transmission cost and privacy concerns. On the other hand, although we can always carry out mining activities at each distributed site, the biased view of the data collected at each site often leads to biased decisions or models, just like the elephant and blind men case. Large random matrices provide natural models for data representations in this context. We can form larger matrices using data matrices from distributed sources. The fundamental mathematical structure (random matrix) is kept invariant under the data fusion. The scalability, however, is relevant.

We can use the unifying tool of random matrix theory to study the resultant problem. The possibility of calculating eigenvalues without explicitly forming the sample covariance matrix allows us to study the problem in a distributed manner. See Section 16.3 for details.

Distributed estimation and detection is natural in this context. See Section 16.1 for details.

Model mining and correlations are the key steps. When the data is independent, identically distributed (i.i.d.)—noise only, the eigenvalue distribution has a rotational symmetry on the complex plane. When signal plus noise is present, some correlations are identified on the complex plane. Non-Hermitian random matrices are studied. This theory is a very recent breakthrough (Chapter 6). □

Example 1.2.5 (mining from sparse, uncertain, and incomplete data) Sparse, uncertain, and incomplete data are defining features for big data applications. For most machine-learning and data-mining algorithms, high-dimensional sparse data cause the reliability of the models derived from the data to deteriorate significantly. We must emphasize that sparsity and high dimensionality are two blessings, rather than curses, for data processing. The concentration of measurement phenomenon—unique to big data—can be exploited [40].

Uncertain data are a special type of data reality where each data field is no longer deterministic but is subject to some random/error distributions. This is mainly linked to domain-specific applications with inaccurate data readings and collection. In this book we promote the exploitation of randomness. Randomness is introduced as a natural resource for our use.

"Incomplete data" refers to missing data field values for some samples. The missing values can be caused by different factors, such as the malfunction of a sensor node, or some systematic policies to intentionally skip some values (e.g., dropping some sensor node readings to save power for transmission). Low-rank matrix recovery [40] deals with incomplete data. Again low-rank matrix recovery takes advantage of high-dimensionality of the data, by using large random matrices as the "sampling" matrix. □

Example 1.2.6 (mining complex and dynamic data) The rise of big data is driven by the rapid increasing of complex data and their changes in volumes and in nature [48]. Documents posted on WWW servers, Internet backbones, social networks, communication networks, and transportation networks, and so on are all featured with complex data. Simple data representations are insufficient. In big data, data types include structured data, unstructured data, semistructured data, and so on. Currently, there is no acknowledged effective and efficient data model to handle big data. In this book we pursue a paradigm of using large random matrices for data representations. This framework has the advantage of uncovering complex relationship networks in data. □

1.3 A Mathematical Introduction to Big Data

There is no standard definition for big data. We give a mathematical definition below.

Definition 1.3.1 (Fundamental Definition for Big Data) Big data must satisfy the following three conditions:

1. Data samples are modeled as random variables, say $X_1, X_2, \ldots, X_n$.

2. The number of data samples, say n, is sufficiently large that some limit results may be observed.
3. A function $f\left(X_1, \ldots, X_n\right)$ can be defined using n random variables.

The main motivation for this definition is to capture the mathematical implications of big data. In particular, we are interested in representing all the data samples in terms of a large random matrix $\mathbf{X}$; applications are modeled as the function $f(\mathbf{X})$.

Example 1.3.2 (data samples are independent random variables) In Definition 1.3.1, most of the time, we consider the special case of Condition 1 when the data samples are modeled by *independent* random variables. Combining Condition 1 with Condition 2, we can take advantage of a very large body of knowledge related to limit theorems in probability and statistics. Roughly speaking, when the size of independent random variables becomes large, some limits are approached.

The simplest and most thoroughly studied example is the sum of independent real-valued random variables. The key to the study of this case is summarized by the trivial but fundamental additive formulas

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{Var}\left(X_i\right)$$

and

$$\psi_{\sum_{i=1}^{n} X_i}(\lambda) = \sum_{i=1}^{n} \psi_{X_i}(\lambda) \tag{1.1}$$

where $\psi_Y(\lambda) = \log \mathbb{E}e^{\lambda Y}$ denotes the logarithm of the moment-generating function of the random variable Y. $\mathbb{E}$ denotes the expectation. These formulas allow one to derive concentration inequalities of $Z = X_1 + X_2 + \cdots + X_n$ around its expectation via Markov's inequality. See [49].

If $X_1, \ldots, X_n$ are independent random variables taking values in $[a_1, b_1], \ldots, [a_n, b_n]$, the additivity formula (1.1) implies that

$$\psi_{Z-\mathbb{E}Z}(\lambda) \leqslant \frac{1}{2}\lambda^2 \nu \text{ for } \lambda \in \mathbb{R}$$

where $\nu = \sum_{i=1}^{n} \left(b_i - a_i\right)^2/4$. Since the right-hand side corresponds to the log-moment generating function of a centered normal random variable with variance ν, $Z - \mathbb{E}Z$ is said to be *sub-Gaussian* with variance factor ν. The sub-Gaussian property implies that $Z - \mathbb{E}Z$ has a sub-Gaussian tail. More precisely, we have, for all $t > 0$

$$\mathbb{P}\left\{[Z - \mathbb{E}Z] \geqslant t\right\} \leqslant 2\exp\left(-t^2/(2\nu)\right)$$

This is Hoeffding's inequality.

One of the simplest and more natural smoothness assumptions that one may consider is the so-called *bounded difference condition*. A function $f : \mathcal{X}^n \to \mathbb{R}$ of n variables

(all taking values in some measurable set $\mathcal{X}$) is said to satisfy the bounded differences condition if constants $c_1, \dots, c_n > 0$ exist such that for every $\mathrm{x}_1, \dots, \mathrm{x}_n, \mathrm{y}_1, \dots, \mathrm{y}_n \in \mathcal{X}^n$

$$\left| f\left(\mathrm{x}_1, \dots, \mathrm{x}_i, \dots, \mathrm{x}_n\right) - f\left(\mathrm{x}_1, \dots, \mathrm{x}_{i-1}, y_i, \mathrm{x}_{i+1}, \dots, \mathrm{x}_n\right) \right| \leqslant c_i$$

In other words, changing any of the n variables, while keeping the rest fixed, cannot cause a big change in the value of the function. Equivalently, one can interpret this as a Lipschitz condition.

The sum of bounded variables is the simplest example of a function of bounded differences. Indeed, if $X_1, \dots, X_n$ are real-valued independent random variables such that X_i takes its values in the interval $[a_i, b_i]$, then $f\left(X_1, \dots, X_n\right) = \sum_{i=1}^{n} X_i$ satisfies the bounded difference condition with $c_i = b_i - a_i$. The basic argument behind the martingale-based approach is that once the function satisfies the bounded difference condition, $Z = f\left(X_1, \dots, X_n\right)$ may be interpreted as a martingale with bounded increments with respect to Doob's filtration. In other words, we may write

$$Z - \mathbb{E}Z = \sum_{i=1}^{n} \Delta_i \tag{1.2}$$

where

$$\Delta_i = \mathbb{E}\left[Z \,|X_1, \dots, X_i\right] - \mathbb{E}\left[Z \,|X_1, \dots, X_{i-1}\right], \quad i = 1, \dots, n$$
$$\Delta_1 = \mathbb{E}\left[Z \,|X_1\right] - \mathbb{E}\left[Z\right].$$

The bounded difference condition implies that, conditionally on $X_1, \dots, X_{i-1}$, the martingale increment Δ_i takes it values in an interval of length at most c_i. Hence, Hoeffding's inequality remains valid for Z with $\nu = (1/4)\sum_{i=1}^{n} c_i^2$. This result is known as the bounded difference inequality, also often called *McDiarmid's inequality*. □

Example 1.3.3 (concentration inequalities for a nonasymptotic theory of independence) The study of random fluctuations of functions of *independent* random variables is the topic of concentration inequalities. Concentration inequalities quantify such statements, typically by bounding the probability that such a function is different from its expected value (or from its median) by more than a certain amount.

In the mid-1990s Michel Talagrand [50] provided major new insight: "a random variable that smoothly depends on the influence of many independent random variables satisfies Chernoff type bounds."

What kind of smooth conditions should we put on a function $f(\cdot)$ of independent random variables $X_1, \dots, X_n$ in order to get concentration bounds for $Z = f\left(X_1, \dots, X_n\right)$ around its mean or median?

One approach to understanding the concentration properties of Lipschitz functions of independent variables is based on investigating how product measures concentrate in *high-dimensional* spaces. The main ideas behind this approach are dominant in Talagrand's work.

In the above examples, we had only considered the linear combination $X_1, \ldots, X_n$ of independent random variables. Now we consider more general combinations $f(\mathbf{X})$ where we write $\mathbf{X} = (X_1, \ldots, X_n)$ for short.

The most powerful concentration of measure results, though, do not just exploit Lipschitz-type behavior in each individual random variable, but *joint* Lipschitz behavior.

One consequence of Talagrand's concentration theorem (Theorem 1.3.5) is the concentration of (empirical) spectral measure for a large random matrix [40].

We say the function $f : \mathbb{C}^n \to \mathbb{R}$ is a 1-Lipschitz function if $|f(\mathbf{x}) - f(\mathbf{y})| \leqslant \|\mathbf{x} - \mathbf{y}\|$ for all random vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, where $\|\cdot\|$ is the Euclidean norm.

Theorem 1.3.4 (Gaussian concentration inequality for Lipschitz functions) Let $X_1, \ldots, X_n \equiv \mathcal{N}(0,1)$ be i.i.d. real Gaussian variables, and let $f : \mathbb{C}^n \to \mathbb{R}$ be a 1-Lipschitz function. Then for any t one has

$$\mathbb{P}\left(|f(\mathbf{X}) - \mathbb{E}f(\mathbf{X})| \geqslant tK\right) \leqslant C\exp\left(-ct^2\right)$$

for some absolute constants $C, c > 0$.

The theorem is valid for all Lipschitz functions for Gaussian random vectors.

Theorem 1.3.5 (Talagrand concentration inequality) Let $K > 0$, and let $X_1, \ldots, X_n$ be independent complex variables with $|X_i| \leqslant K$ for all $1 \leq i \leq n$. Let $f : \mathbb{C}^n \to \mathbb{R}$ be a 1-Lipschitz and convex function. Then for any t one has

$$\mathbb{P}\left(|f(\mathbf{X}) - \mathbb{M}f(\mathbf{X})| \geqslant tK\right) \leqslant C\exp\left(-ct^2\right)$$
$$\mathbb{P}\left(|f(\mathbf{X}) - \mathbb{E}f(\mathbf{X})| \geqslant tK\right) \leqslant C\exp\left(-ct^2\right)$$

for some absolute constants $C, c > 0$, where $\mathbb{M}f(\mathbf{X})$ is a median of $f(\mathbf{X})$

The theorem is valid for all Lipschitz and convex functions for independent (not necessarily Gaussian) random vectors.

Example 1.3.6 (large random matrix theory) Random matrix theory or quantum information theory is very relevant for big data. The vision of exploiting random matrixes to model big data is explicitly proposed in [39].

For N random (row) vectors $\mathbf{X}_1, \ldots, \mathbf{X}_N \in \mathbb{C}^{1\times T}$, we form an $N \times T$ random matrix

$$\mathbf{X} = \left(\mathbf{X}_1, \ldots, \mathbf{X}_N\right)^T \in \mathbb{C}^{N\times T}$$

We say a matrix $\mathbf{Y}$ is Hermitian if $\mathbf{Y} = \mathbf{Y}^H$, where H denotes the conjugate and transpose of a matrix. In general, the random matrix $\mathbf{X}$ is not Hermitian.

The classical framework is to study the regime of N fixed while $T \to \infty$. For modern big data, this fundamental assumption is invalid. We must study the new paradigm

$$N \to \infty, T \to \infty \text{ but N/T} \to c \in [0, \infty)$$

where c is a fixed constant.

This book surveys a lot of recent results of the literature in Part I. □

Example 1.3.7 (free probability theory for hermitian random matrices) When the sizes of random matrices are large, conventional independence is replaced with asymptotically freeness. Free random variables may be thought of as "independent" random matrices in the classical sense. Chapter 5 applies this theory to model large random matrices. Free random variables are random infinite-dimensional linear operators that are equivalently very large random matrices. The statistical properties of free random variables are equivalently those of the eigenvalues of large random matrices. □

Free probability theory was introduced by Voiculescu around 1983 in order to attack the isomorphism problem of von Neumann algebras of free groups. Voiculescu isolated a structure showing up in this context, which he named "freeness." His fundamental insight was to separate this concept from its operator algebraic origin and investigate it for its own sake. Furthermore, he promoted the point of view that freeness should be seen as a noncommutative analog of the classical probabilistic concept of "independence" for random variables. Hence freeness is also called "free independence" and the whole subject became known as "free probability theory."

The theory was lifted to a new level when Voiculescu discovered, in 1991, that the freeness property is also present for many classes of *random matrices*, in the asymptotic regime when the size of the matrices tends to infinity. This insight, bringing together the a priori entirely different theories of operator algebras and of random matrices, had quite some impact in both directions. Modeling operator algebras by random matrices resulted in some insightful results about operator algebras, whereas tools developed in operator algebras and free probability theory could now be applied to random matrix problems, yielding, in particular, new ways to calculate the asymptotic eigenvalue distribution of many random matrices. Freeness is motivated not by its initial occurrence in operator algebras but by its random matrix connection.

In free probability theory, the central limit theorem on the sum of independent free random variables gives a semicircle distribution. A semicircle distribution serves the same function as the Gaussian or normal distribution for the sum of independent commuting random variables. If $X_1, X_2, \ldots, X_n$ are identically distributed zero mean free random variables with variance of $(R/2)^2$, the free summation or additive free convolution of

$$\frac{1}{\sqrt{n}} X_1 \boxplus X_2 \boxplus \cdots \boxplus X_n$$

has the semicircle distribution of

$$p(t) = \begin{cases} \frac{1}{2\pi R^2}\sqrt{R^2 - t^2} & |t| \leqslant R \\ 0 & \text{otherwise,} \end{cases}$$

where R is the radius of the distribution and $\boxplus$ denotes the additive free convolution.

Using free probability, we can calculate the histogram for a generic realization of a 3000×3000 random matrix $p(X, Y)$, where X and Y are, respectively, independent Gaussian and Wishart random matrices: $p(X, Y) = X + Y$; $P(X, Y) = XY + YX + X^2$. $P(X, Y)$ is a polynomial of two random matrices.

Example 1.3.8 (free probability theory for non-Hermitian random matrices) As pointed out above, in general, a random matrix $\mathbf{Y}$ is non-Hermitian. Most tools in algebra deal with Hermitian random matrices. Non-Hermitian random matrices are much more difficult to handle, compared with their Hermitian counterparts. Chapter 6 gives a comprehensive introduction to model the data using (large) non-Hermitian random matrices. □

The eigenvalue density of a product

$$\mathbf{X}_1\mathbf{X}_2\cdots\mathbf{X}_L \tag{1.3}$$

of $L \geq 2$ independent $N \times N$ Gaussian random matrices in the limit $N \to \infty$ is rotationally symmetric in the complex plane and is given by a simple expression

$$\rho\left(z,\overline{z}\right) = \begin{cases} \frac{1}{L\pi}\sigma^{-2/L}|z|^{-2+2/L} & |z| \leqslant \sigma \\ 0 & |z| > \sigma \end{cases}$$

where the $\overline{z}$ denotes the complex conjugate of a complex number z, and the effective scale parameter $\sigma = \sigma_1\sigma_2\cdots\sigma_L$. We have

$$\mathbb{E}\left(\mathbf{X}_1\right)_{ij} = \cdots = \mathbb{E}\left(\mathbf{X}_L\right)_{ij} = 0, \quad i,j = 1,\ldots,N$$
$$\mathbb{E}\left|\left(\mathbf{X}_1\right)_{ij}\right|^2 = \sigma_1^2/N, \cdots, \mathbb{E}\left|\left(\mathbf{X}_L\right)_{ij}\right|^2 = \sigma_L^2/N, \quad i,j = 1,\ldots,N$$

The parameter σ corresponds to the radius of the circular support and is related to the amplitude of the Gaussian fluctuations. This form of the eigenvalue density is universal. It is identical for products of Gaussian Hermitian, non-Hermitian, real or complex random matrices. It does not change even if the matrices in the product are taken from different Gaussian ensembles.

Study the product

$$\mathbf{P} = \mathbf{A}_1\mathbf{A}_2\cdots\mathbf{A}_L \tag{1.4}$$

of $L \geq 1$ independent rectangular large random Gaussian matrices $\mathbf{A}_l, l = 1,2,\ldots,L$ of dimensions $N_l \times N_{l+1}$. We are interested in the limit $N_{L+1} \to \infty$ and

$$R_l \equiv \frac{N_l}{N_{l+1}} = \text{finite}, \quad \text{for } l = 1,2,\ldots,L+1$$

The σ_l parameters set the scale for the Gaussian fluctuations in $\mathbf{A}_l$s. The entries of each matrix $\mathbf{A}_l$ can be viewed as independent centered Gaussian random variables, the variance of the real and imaginary parts being proportional to σ_l^2 and inversely proportional to the square root of the number N_lN_{l+1} of elements in the matrix.

Consider

$$\mathbf{Q} = \mathbf{P}^H\mathbf{P}, \qquad \mathbf{R} = \mathbf{P}\mathbf{P}^H$$

where $\mathbf{P}$ is defined in (1.4). $\mathbf{Q}$ and $\mathbf{R}$ are are Hermitian, and they have non-negative spectra, which differ only in the zero modes. The M transform of the matrix $\mathbf{X}$ is defined as

$$M_{\mathbf{X}}\left(z,\overline{z}\right)=zG_{\mathbf{X}}\left(z,\overline{z}\right)-1$$

where $G_{\mathbf{X}}\left(z,\overline{z}\right)$ is the Green's function.

The main finding is that the eigenvalue distribution and the M transform of the product [(1.4)] are spherically symmetric. We shall show the M transform to satisfy the L-th order polynomial equation:

$$\prod_{l=1}^{L}\left(\frac{M_{\mathbf{P}}\left(|z|^2\right)}{R_l}+1\right)=\frac{|z|^2}{\sigma^2} \tag{1.5}$$

where the scale parameter is $\sigma=\sigma_1\sigma_2\cdots\sigma_M$.

An analogous equation for $\mathbf{Q}$ reads

$$\sqrt{R_l}\frac{M_{\mathbf{Q}}(z)+1}{M_{\mathbf{Q}}(z)}\prod_{l=1}^{L}\left(\frac{M_{\mathbf{Q}}(z)}{R_l}+1\right)=\frac{z}{\sigma^2} \tag{1.6}$$

The free argument in (1.5) is $|z|^2$, and z in (1.6). It is surprising that there is rotational symmetry in the complex plane for the product of Gaussian random matrices $\mathbf{P}$, while the study of the Hermitian product $\mathbf{Q}$ breaks the rotational symmetry. In other words, given a data matrix $\mathbf{A}_l, l=1,\ldots,L$, some statistical structure (symmetry) will be lost if we study the non-negative Hermitian random matrix $\mathbf{Q}$, instead of non-Hermitian random matrix $\mathbf{P}$.

One unexpected implication of the universality is that a product of random matrices whose spectra do not necessarily display rotational symmetry has an eigenvalue distribution that does possess rotational symmetry on the complex plane (i.e., the average density depends only on $|\lambda|$).

A random quantum state is defined by specifying a probability measure in the space of density matrices $\boldsymbol{\rho}$, i.e., Hermitian, weakly positive-definite (i.e., with nonnegative eigenvalues), and normalized (i.e., $\mathrm{Tr}\,\boldsymbol{\rho}=1$) matrices. For any rectangular matrix $\mathbf{Z}$, one can define $\boldsymbol{\rho}\equiv\mathbf{Z}\mathbf{Z}^H/\mathrm{Tr}\left(\mathbf{Z}\mathbf{Z}^H\right)$ is a proper random quantum density matrix.

If we model the system using the random states through $\boldsymbol{\rho}$, we will break the rotational symmetry of the product of random matrices $\mathbf{P}$ (defined in (1.4)) in the complex plane. It makes sense because the eigenvalues of $\mathbf{P}$ are distributed in the complex plane and the eigenvalues of $\boldsymbol{\rho}$ are in the real axis (non-negative real values).

In statistics we often use a sample covariance matrix in the form of $\mathbf{Q}$. The comments for $\boldsymbol{\rho}$ are also valid for a sample covariance matrix. By studying the sample covariance matrix we will lose some structure information (such as rotational symmetry in the complex plane for the M transform). See Example 1.3.9 for the potential relevance to applications.

We now consider the eigenvalue statistics for complex $N\times N$ Wishart matrices $\mathbf{X}_{r,s}^H\mathbf{X}_{r,s}$, where $\mathbf{X}_{r,s}$ is equal to the product of r complex Gaussian matrices, and the inverse of s complex Gaussian matrices. In particular, we have

$$\mathbf{X}_{r,s}=\mathbf{G}_r\mathbf{G}_{r-1}\cdots\mathbf{G}_1\left(\tilde{\mathbf{G}}_s\tilde{\mathbf{G}}_{s-1}\cdots\tilde{\mathbf{G}}_1\right)^{-1}$$

where each $\mathbf{G}_k$ is a rectangular standard complex Gaussian matrix of dimension $n_k \times n_{k-1}, n_k \geqslant n_{k-1},$ and $n_0 = N$, and each $\tilde{\mathbf{G}}_k$ is a square of dimension $N \times N$.

Example 1.3.9 (functional averages over Gaussian ensembles) The MIMO channel model is defined similarly to (3.11). The result here can be applied to massive MIMO analysis. See Section 15.3. We repeat the definition to fix a different notation. Denoting the number of transmitting antennas by M and the number of receiving antennas by N, the channel model is

$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{n} \tag{1.7}$$

where $\mathbf{s} \in \mathbb{C}^M$ is the transmitted vector, $\mathbf{y} \in \mathbb{C}$ is the received vector, $\mathbf{H} \in \mathbb{C}^{N \times M}$ is a complex matrix and $\mathbf{n} \in \mathbb{C}^N$ is the zero mean complex Gaussian vector with independent, equal variance entries. We assume that $\mathbb{E}\left(\mathbf{n}\mathbf{n}^H\right) = \mathbf{I}_N$, where $(\cdot)^H$ denotes the complex conjugate transpose and $\mathbf{I}_N$ the $N \times N$ identity matrix. It is reasonable to put a power constraint

$$\mathbb{E}\left(\mathbf{n}^H\mathbf{n}\right) = \mathbb{E}\left[\mathrm{Tr}\left(\mathbf{n}\mathbf{n}^H\right)\right] \leqslant P$$

where P is the total transmitted power. The signal-to-noise ratio, denoted by snr, is defined as the quotient of the signal power and the noise power, and in this case is equal to P/N. □

Recall that if $\mathbf{A}$ is an $n \times n$ Hermitian matrix then there exists $\mathbf{U}$ unitary and $\mathbf{D} = \mathrm{diag}\left(d_1, \ldots, d_n\right)$ such that $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^H$. Given a continuous function f, we define $f(\mathbf{A})$ as

$$f(\mathbf{A}) = \mathbf{U}\,\mathrm{diag}\left(f\left(d_1\right), \ldots, f\left(d_n\right)\right)\mathbf{U}^H$$

Naturally, the simplest example is the one where $\mathbf{H}$ has independent and identically distributed (i.i.d.) Gaussian entries, which constitutes the canonical model for the single-user narrow band MIMO channel. It is known that the capacity of this channel is achieved when $\mathbf{s}$ is a vector with complex Gaussian zero mean and covariance $\mathrm{snr}\,\mathbf{I}_M$. See [51, 52] for instance. For the fast fading channel, assuming statistical channel state information at the transmitter, the ergodic capacity is given by

$$\mathbb{E}\left[\log\det\left(\mathbf{I}_N + \mathrm{snr}\,\mathbf{H}\mathbf{H}^H\right)\right] = \mathbb{E}\left[\mathrm{Tr}\log\left(\mathbf{I}_N + \mathrm{snr}\,\mathbf{H}\mathbf{H}^H\right)\right] \tag{1.8}$$

where in the last equality we use the fundamental fact that

$$\log\det(\cdot) = \mathrm{Tr}\log(\cdot) \tag{1.9}$$

We prefer the form of $\mathrm{Tr}\log(\cdot)$ because the trace $\mathrm{Tr}(\cdot)$ is a linear function. The expectation $\mathbb{E}(\cdot)$ is also a linear function. Sometimes it is convenient to exchange the order of $\mathbb{E}$ and $\mathrm{Tr}(\cdot)$ in (1.8):

$$\begin{aligned}\mathbb{E}\left[\log\det\left(\mathbf{I}_N + \mathrm{snr}\,\mathbf{H}\mathbf{H}^H\right)\right] &= \mathbb{E}\left[\mathrm{Tr}\log\left(\mathbf{I}_N + \mathrm{snr}\,\mathbf{H}\mathbf{H}^H\right)\right] \\ &= \mathrm{Tr}\left[\mathbb{E}\log\left(\mathbf{I}_N + \mathrm{snr}\,\mathbf{H}\mathbf{H}^H\right)\right]\end{aligned}$$

The $\mathbb{E}(\mathbf{X})$ can be approximated by the arithmetic average $\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i$ when n "snapshots" of the $p \times p$ random matrix $\mathbf{X}$ are observed. As a result, we reach

$$\begin{aligned}\mathbb{E}\left[\log\det\left(\mathbf{I}_N + \text{snr}\,\mathbf{H}\mathbf{H}^H\right)\right] &= \mathbb{E}\left[\text{Tr}\log\left(\mathbf{I}_N + \text{snr}\,\mathbf{H}\mathbf{H}^H\right)\right] \\ &= \text{Tr}\left[\mathbb{E}\log\left(\mathbf{I}_N + \text{snr}\,\mathbf{H}\mathbf{H}^H\right)\right] \\ &\approx \frac{1}{n}\,\text{Tr}\left[\sum_{i=1}^{n}\log\left(\mathbf{I}_N + \text{snr}\,\mathbf{H}_i\mathbf{H}_i^H\right)\right] \end{aligned} \tag{1.10}$$

which boils down to the sum of random positive definite Hermitian matrices $\mathbf{H}_i\mathbf{H}_i^H$, $i = 1, \ldots, n$, given the i-th "snapshot" $\mathbf{H}_i$ of the random channel matrix $\mathbf{H}$ that is defined in (3.16). See [40] for a whole chapter on the sum of random matrices. The channel capacity with a finite number of samples can be obtained using (1.10). Note that the Frobenius norm is defined as

$$\|\mathbf{B}\|_F^2 \equiv \text{Tr}\left(\mathbf{B}\mathbf{B}^H\right)$$

In (1.10), if we expand the function $\log\left(\mathbf{I}_N + \text{snr}\,\mathbf{H}_i\mathbf{H}_i^H\right)$ using its Taylor series, we can reduce the problem to the sample moments m_k defined as

$$\hat{m}_k = \frac{1}{M}\,\text{Tr}\left[\left(\frac{1}{N}\mathbf{H}_i\mathbf{H}_i^H\right)^k\right]$$

for an integer $k \geq 1$. Because the sample moments $\hat{m}_k$ are *consistent estimators* of true moments m_k, it is then natural to use the moment method for the inference of the parameters [53, p. 425]. See Section 8.9.3 for this connection.

More generally, we can expand a functional of a random matrix in the form of $f\left(\mathbf{H}\mathbf{H}^H\right)$ in terms of its Taylor series. We can similarly obtain the true moments m_k. We can use sample moments $\hat{m}_k$ to estimate the true moments.

Another important performance measure is the minimum mean square error (MMSE) achieved by a linear receiver, which determines the maximum achievable output signal to interference and noise ratio (SINR). For an input vector $\mathbf{x}$ with i.i.d. entries of zero mean and unit variance, the MSE at the output of the MMSE receiver is given by

$$\min_{\mathbf{M}\in\mathbb{C}^{M\times N}}\mathbb{E}\left[\|\mathbf{x} - \mathbf{M}\mathbf{y}\|^2\right] = \mathbb{E}\left[\text{Tr}\log\left(\mathbf{I}_M + \text{snr}\,\mathbf{H}^H\mathbf{H}\right)^{-1}\right] \tag{1.11}$$

where the expectation on the left-hand side is over both the vectors x and the random matrices $\mathbf{H}$, whereas the right-hand side is over $\mathbf{H}$ only. See [52] for details.

Let $\mathbf{H}$ be an $n \times n$ Gaussian random matrix with complex, independent, and identically distributed entries of zero mean and unit variance. Given an $n \times n$ positive definite matrix $\mathbf{A}$, and a continuous function $f : \mathbb{R}^+ \to \mathbb{R}$ such that $\int_0^\infty e^{-\alpha t}|f(t)|^2 dt < \infty$ for every $\alpha > 0$, Tucci and Vega (2013) [54] find a new formula for the expectation

$$\mathbb{E}\left[\text{Tr}\left(f\left(\mathbf{H}\mathbf{A}\mathbf{H}^H\right)\right)\right]$$

Taking $f(x) = \log(1 + x)$ gives another formula for the capacity of the MIMO communication channel, and taking $f(x) = (1 + x)^{-1}$ gives the MMSE achieved by a linear receiver.

From Example 1.3.8, we see the connection of eigenvalues of $\mathbf{H}$ and $\mathbf{H}\mathbf{H}^H$, when $\mathbf{H}$ is decomposed into a product of L random matrices.

Example 1.3.10 (matrix hypothesis testing) Applications include: (i) anomaly detection; (ii) denial of service for big data; (iii) bad data detection for Smart Grid (state estimation). We consider the so-called matrix hypothesis-testing problem

$$\begin{aligned} \mathcal{H}_0 &: \quad \mathbf{Y} = \mathbf{X} \\ \mathcal{H}_1 &: \quad \mathbf{Y} = \sqrt{\text{SNR}} \cdot \mathbf{H} + \mathbf{X} \end{aligned} \tag{1.12}$$

where SNR represents the signal-to-noise ratio, and $\mathbf{X}$ is a non-Hermitian random matrix of $m \times n$. We further assume that $\mathbf{H}$ is independent of $\mathbf{X}$. The problem of (1.12) is equivalent to

$$\begin{aligned} \mathcal{H}_0 &: \quad \mathbf{Y}\mathbf{Y}^H = \mathbf{X}\mathbf{X}^H \\ \mathcal{H}_1 &: \quad \mathbf{Y}\mathbf{Y}^H = \text{SNR} \cdot \mathbf{H}\mathbf{H}^H + \mathbf{X}\mathbf{X}^H + \sqrt{\text{SNR}}\left(\mathbf{H}\mathbf{X}^H + \mathbf{X}\mathbf{H}^H\right) \end{aligned} \tag{1.13}$$

where $\mathbf{H}\mathbf{H}^H, \mathbf{X}\mathbf{X}^H, \mathbf{Y}\mathbf{Y}^H$ are positive semidefinite Hermitian random matrices, which are Wishart matrices if $\mathbf{X}, H$ are Gaussian random matrices. A matrix $\mathbf{A}$ of $m \times n$ is said to be positive semidefinite if all the eigenvalues of $\mathbf{A}$ are non-negative, i.e., $\lambda_i(\mathbf{A}) \geq 0, i = 1, \ldots, \min(m, n)$. The matrix $\left(\mathbf{H}\mathbf{X}^H + \mathbf{X}\mathbf{H}^H\right)$ is Hermitian. □

The likelihood ratio test (LRT) is the natural choice. We deal with matrix-valued random variables, where the matrix sizes are large. See Section 8.11 for details. The analysis of these metrics requires advanced tools, such as the nonasymptotic theory of random matrices. The nonasymptotic theory is based on the "concentration of measure" phenomenon when the size of a matrix is large but finite. This phenomenon is the starting point for almost all the results.

Theorem 17.3.1 essentially says that if we take two large random matrices $\mathbf{A}_N$ and $\mathbf{B}_N$, and if we conjugate one of them by a uniformly random unitary transformation $\mathbf{U}_N$, then the resulting pair of matrices $\mathbf{A}_N$ and $\mathbf{U}_N\mathbf{B}_N\mathbf{U}_N^H$ will be approximately free. As a slogan, this can be expressed as follows

Two large random matrices
in the general position
are asymptotically free!

For a multivariate Gaussian distribution $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, it is well known that the differential entropy $H(\cdot)$ is given by

$$\mathcal{H}(\boldsymbol{\Sigma}) = \frac{p}{2} + \frac{1}{2}p\log(2\pi) + \frac{1}{2}\log\det\boldsymbol{\Sigma}. \tag{1.14}$$

The high-dimensional setting where the dimension $p(n)$ grows with the sample size n is of particular current interest.

Let $\mathbf{X}_1, \ldots, \mathbf{X}_{n+1}$ be an independent random sample from the p-dimensional Gaussian distribution $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The sample covariance matrix is

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n+1}\left(\mathbf{X}_k - \overline{\mathbf{X}}\right)\left(\mathbf{X}_k - \overline{\mathbf{X}}\right)^T$$

A central limit theorem is established for the log determinant of $\hat{\mathbf{\Sigma}}$ in the high-dimensional setting where the dimension p grows with the sample size n with the only restriction that $p(n) \leq n$. In the case when $\lim_{n\to 0} \frac{p(n)}{n} = r$ for some $0 \leq r \leq 1$, the central limit theorem shows

$$\frac{\log\det\hat{\mathbf{\Sigma}} - \sum_{k=1}^{p} \log\left(1 - \frac{k}{n}\right) - \log\det\mathbf{\Sigma}}{\sqrt{-2\log\left(1 - \frac{p}{n}\right)}} \xrightarrow{Law} \mathcal{N}(0,1) \text{ as } n \to \infty \tag{1.15}$$

The result for the boundary case $p = n$ yields

$$\frac{\log\det\hat{\mathbf{\Sigma}} - \log(n-1)! + n\log n - \log\det\mathbf{\Sigma}}{\sqrt{2\log n}} \xrightarrow{Law} \mathcal{N}(0,1) \text{ as } n \to \infty \tag{1.16}$$

One common problem in statistics and engineering is to estimate the distance between two population distributions based on the samples. A commonly used measure of closeness is the relative entropy or the Kullback–Leibler divergence. For two distributions $\mathbb{P}$ and $\mathbb{Q}$ with respective density functions $p(\cdot)$ and $q(\cdot)$, the relative entropy between $\mathbb{P}$ and $\mathbb{Q}$ is

$$KL(\mathbb{P},\mathbb{Q}) = \int p(x)\log\frac{p(x)}{q(x)}dx$$

In the case of two multivariate Gaussian distributions $\mathbb{P} = \mathcal{N}_p\left(\boldsymbol{\mu}_1, \mathbf{\Sigma}_1\right)$, $\mathbb{Q} = \mathcal{N}_p\left(\boldsymbol{\mu}_2, \mathbf{\Sigma}_2\right)$

$$2KL(\mathbb{P},\mathbb{Q}) = \mathrm{Tr}\left(\mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1\right) - p + \left(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\right)^T \mathbf{\Sigma}_2^{-1}\left(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\right) + \log\left(\frac{\det\mathbf{\Sigma}_1}{\det\mathbf{\Sigma}_2}\right) \tag{1.17}$$

From (1.17), it is clear that estimation of the relative entropy involves estimation of the log determinants $\log\det\mathbf{\Sigma}_1$ and $\log\det\mathbf{\Sigma}_2$.

For testing the hypothesis that two multivariate Gaussian distributions $\mathbb{P} = \mathcal{N}_p\left(\boldsymbol{\mu}_1, \mathbf{\Sigma}_1\right)$, and $\mathbb{Q} = \mathcal{N}_p\left(\boldsymbol{\mu}_2, \mathbf{\Sigma}_2\right)$ have the same entropy, we have

$$\mathcal{H}_0 : \mathcal{H}(\mathbb{P}) = \mathcal{H}(\mathbb{Q}) \text{ versus } \mathcal{H}_1 : \mathcal{H}(\mathbb{P}) \neq \mathcal{H}(\mathbb{Q})$$

For any given significance level $0 < \alpha < 1$, a test with the asymptotic level α can be constructed easily using the central limit theorem given above, based on two independent samples, one from $\mathbb{P}$ and another from $\mathbb{Q}$.

Knowledge of the log determinant of covariance matrices is also essential for the quadratic discriminant analysis (QDA). For classification of two multivariate Gaussian distributions $\mathcal{N}_p\left(\boldsymbol{\mu}_1, \mathbf{\Sigma}_1\right)$ and $\mathcal{N}_p\left(\boldsymbol{\mu}_2, \mathbf{\Sigma}_2\right)$, when the parameters $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2$ are known, the oracle discriminant is

$$\Delta = -\left(\mathbf{z} - \boldsymbol{\mu}_1\right)^T \mathbf{\Sigma}_1^{-1}\left(\mathbf{z} - \boldsymbol{\mu}_1\right) + \left(\mathbf{z} - \boldsymbol{\mu}_2\right)^T \mathbf{\Sigma}_2^{-1}\left(\mathbf{z} - \boldsymbol{\mu}_2\right) - \log\left(\frac{\det\mathbf{\Sigma}_1}{\det\mathbf{\Sigma}_2}\right) \tag{1.18}$$

That is, the observation vector z is classified into the population with $\mathcal{N}_p\left(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right)$ distribution if $\Delta > 0$ and into $\mathcal{N}_p\left(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2\right)$ otherwise.

Example 1.3.11 (outliers in signal plus noise) For a complex variable $z = x + iy$, the Dirac delta function is defined by $\delta^2(z) \equiv \delta(x)\,\delta\left(y\right)$, and we define $\partial/\partial\overline{z} = \left(\partial/\partial x + i\partial/\partial y\right)/2$, and $\partial/\partial z = \left(\partial/\partial x - i\partial/\partial y\right)/2$. For simplicity, we use the notation $f(z)$ (instead of $f\left(z, \overline{z}\right)$) for general, nonholomorphic functions on the complex plane. □

We provide a general formula for the eigenvalue density of large random $N \times N$ matrices of the form

$$\mathbf{A} = \mathbf{M} + \mathbf{LXR} \tag{1.19}$$

where $\mathbf{M}$, $\mathbf{L}$ and $\mathbf{R}$ are general ($\mathbf{M}$) or arbitrary invertible ($\mathbf{L}$ and $\mathbf{R}$) deterministic matrices, and $\mathbf{X}$ is a random matrix of zero-mean independent and identically distributed (i.i.d.) elements with zero mean and variance $1/N$. For example, the entries of $\mathbf{X}$ are Gaussian or Bernoulli random variables. The model (1.19) has been used to model the brain, and may be used for sensor networks and wireless networks.

As $\mathbf{X}$ and therefore $\mathbf{L}XR$ have zero mean, $\mathbf{M}$ is the ensemble average of $\mathbf{A}$. The random fluctuations of $\mathbf{A}$ around its average are given by the matrix $\mathbf{L}XR$, which for general $\mathbf{L}$ and/or $\mathbf{R}$ has dependent and nonidentically distributed elements, due to the possible mixing and nonuniform scaling of the rows (columns) of the i.i.d. $\mathbf{X}$ by $\mathbf{L}$ ($\mathbf{R}$).

The density of the eigenvalues of $\mathbf{A} = \mathbf{M} + \mathbf{LXR}$, in the complex plane for a realization of $\mathbf{X}$ (also known as the empirical spectral distribution), is defined by

$$\rho_{\mathbf{X}}(z) = \frac{1}{N}\sum_{i=1}^{N}\delta^2\left(z - \lambda_i\right)$$

where λ_i are the eigenvalues of $\mathbf{M} + \mathbf{LXR}$. It is known [55] that $\rho_{\mathbf{X}}(z)$ is asymptotically self-averaging, in the sense that with probability 1, $\rho_{\mathbf{X}}(z) - \rho(z)$ converges to zero (in the distributional sense) as $N \to \infty$, where $\rho(z) \equiv \langle \rho_{\mathbf{X}}(z) \rangle_{\mathbf{X}}$ is the ensemble average of $\rho_{\mathbf{X}}(z)$. Thus for large enough N, any typical realization of $\mathbf{X}$ yields an eigenvalue density $\rho_{\mathbf{X}}(z)$ that is arbitrarily close to $\rho(z)$.

For any matrix $\mathbf{B}$, we denote its operator norm (its maximum singular value) by $\|\mathbf{B}\|$ and we define its (normalized) Frobenius norm via

$$\|\mathbf{B}\|_F \equiv \frac{1}{N}\sum_{i,j=1}^{N}\left|B_{ij}\right|^2 = \frac{1}{N}\,\mathrm{Tr}\left(\mathbf{B}\mathbf{B}^H\right) \tag{1.20}$$

(equivalently, $\|\mathbf{B}\|_F$ is the root mean square of the singular values of $\mathbf{B}$).

Our general result is that for large N, $\rho(z)$ is nonzero in the region of complex plane satisfying

$$\frac{1}{N}\,\mathrm{Tr}\left[\left(\mathbf{M}_z\mathbf{M}_z^{\dagger}\right)^{-1}\right] \geqslant 1 \tag{1.21}$$

where we defined

$$\mathbf{M}_z = L^{-1}\left(z\mathbf{I} - \mathbf{M}\right)\mathbf{R}^{-1}$$

Using (1.20) , we can express (1.21) as

$$\left\|\mathbf{R}(z\mathbf{I}-\mathbf{M})^{-1}\mathbf{L}\right\|_F \geqslant 1$$

inside this region, $\rho(z)$ is given by

$$\rho(z) = \frac{1}{N}\frac{1}{z}\frac{\partial}{\partial \overline{z}} \operatorname{Tr}\left[(\mathbf{RL})^{-1}\mathbf{M}_z^H\left(\mathbf{M}_z\mathbf{M}_z^H + g(z)^2\right)^{-1}\right] \tag{1.22}$$

where $g(z)$ is a real, scalar function found by solving

$$\frac{1}{N}\operatorname{Tr}\left[\left(\mathbf{M}_z\mathbf{M}_z^H + g^2\right)^{-1}\right] = 1,$$

for g for each z.

Example 1.3.12 (asymptotically deterministic character of limiting spectral distributions) One motivation is to study the random block matrices. We consider $N \times N$ matrices that are Hermitian with above diagonal "block-rows" (or "strips") of height bounded by a constant d. Examples are matrices with i.i.d block entries but the theory we develop here applies more generally. □

Our analysis is often based on Stieltjes transforms. We call

$$m_n(z) = \frac{1}{N}\operatorname{Tr}\left(\left(\mathbf{M} - z\mathbf{I}_N\right)^{-1}\right)$$

the Stieltjes transform of $\mathbf{M}$, the $N \times N$ random matrix of interest. Here $\mathbf{I}_N$ is the $N \times N$ identity matrix. In much of our analysis, we will let N grow to infinity.

Theorem 1.3.13 Suppose the $N \times N$ Hermitian matrix $\mathbf{M}$ can be written as

$$\mathbf{M} = \sum_{i=1}^{n} \mathbf{M}_i,$$

where $\mathbf{M}_i$ are *independent* with $\operatorname{rank}\left(\mathbf{M}_i\right) \leqslant d_i$. Let $z \in \mathbb{C}^+$ and $\operatorname{Im}[z] = v > 0$. Call

$$m_n(z) = \frac{1}{N}\operatorname{Tr}\left(\left(\mathbf{M} - z\mathbf{I}_N\right)^{-1}\right)$$

Then, for any $t > 0$,

$$\mathbb{P}\left(\left|m_n(z) - \mathbb{E}\left(m_n(z)\right)\right| > t\right) \leqslant C\exp\left(-c\frac{N^2v^2t^2}{\sum_{i=1}^{n} d_i^2}\right)$$

where C and c are two constants that do not depend on n or d_is.

We can extend the above theorem to the following.

Theorem 1.3.14 Suppose the $N \times N$ Hermitian matrix $\mathbf{M}$ can be written as

$$\mathbf{M} = \sum_{1 \leqslant i,j \leqslant n} \mathbf{\Theta}_{i,j}$$

where $\mathbf{\Theta}_{i,j} = f_{i,j}\left(Z_i, Z_j\right)$ is a $N \times N$ matrix and the random variables $\left\{Z_i\right\}_{i=1}^{n}$ are independent. ($f_{i,j}\left(Z_i, Z_j\right)$ are simply matrix valued functions of our random variables.) Let $\mathbf{M}_i$ be the Hermitian matrix

$$\mathbf{M}_i = \mathbf{\Theta}_{i,i} + \sum_{j \neq i} \left(\mathbf{\Theta}_{i,j} + \mathbf{\Theta}_{j,i}\right)$$

Assume that rank $\left(\mathbf{M}_i\right) \leqslant d_i$. Let $z \in \mathbb{C}^+$ and $\operatorname{Im}[z] = v > 0$. Call

$$m_n(z) = \frac{1}{N} \operatorname{Tr}\left(\left(\mathbf{M} - z\mathbf{I}_N\right)^{-1}\right)$$

Then, for any $t > 0$,

$$\mathbb{P}\left(\left|m_n(z) - \mathbb{E}\left(m_n(z)\right)\right| > t\right) \leqslant C \exp\left(-c \frac{N^2 v^2 t^2}{\sum_{i=1}^{n} d_i^2}\right)$$

where C and c are two constants that do not depend on n nor d_is.

The previous theorem is derived from the following theorem.

Theorem 1.3.15 Suppose that the $N \times N$ Hermitian matrix $\mathbf{M}$ is such that, for independent random variables $\left\{Z_i\right\}_{i=1}^{n}$ and a matrix valued function f,

$$\mathbf{M} = f\left(Z_1, \ldots, Z_n\right)$$

Suppose further that for all $1 \leq i \leq n$, there exists a matrix $\mathbf{N}_i$ such that

$$\mathbf{N}_i = f_i\left(Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n\right)$$

and rank $\left(\mathbf{M} - \mathbf{N}_i\right) \leqslant d_i$. (When $i = 1$, $\mathbf{N}_1 = f_1\left(Z_2, \ldots, Z_n\right)$ and when $i = n$, $\mathbf{N}_n = f_n\left(Z_1, \ldots, Z_{n-1}\right)$. f_is are simply matrix-valued functions.) Let $z \in \mathbb{C}^+$ and $\operatorname{Im}[z] = v > 0$. Call

$$m_n(z) = \frac{1}{N} \operatorname{Tr}\left(\left(\mathbf{M} - z\mathbf{I}_N\right)^{-1}\right)$$

Then, for any $t > 0$,

$$\mathbb{P}\left(\left|m_n(z) - \mathbb{E}\left(m_n(z)\right)\right| > t\right) \leqslant C \exp\left(-c \frac{N^2 v^2 t^2}{\sum_{i=1}^{n} d_i^2}\right)$$

where C and c are two constants that do not depend on n nor d_is.

This is McDiarmid-type inequality.

Example 1.3.16 (random particles) Beyond random matrices, how about the empirical measure of random particles in $\mathbb{R}^d$? Is there an analog of the circular law phenomenon? Does the ball replace the disc? The answer is positive. A wireless radio sensor can be modeled as a random particle, for example. □

We consider a system of N particles in $\mathbb{R}^d$ at positions $\mathbf{x}_1, \ldots, \mathbf{x}_N$, say with charge $1/N$. These particles are subject to confinement by an external field via a potential $\mathbf{x} \in \mathbb{R}^d \mapsto V(\mathbf{x})$. and to internal pair interaction (typically repulsion) via a potential $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^d \times \mathbb{R}^d \mapsto W(\mathbf{x}, \mathbf{y})$. The idea is that an equilibrium may emerge as N tends to infinity. The configuration energy is

$$\begin{aligned}\mathcal{I}_N\left(\mathbf{x}_1, \ldots, \mathbf{x}_N\right) &= \frac{1}{N}\sum_{i=1}^{N} V\left(\mathbf{x}_i\right) + \frac{1}{N^2}\sum_{1\leqslant i<j\leqslant N} W\left(\mathbf{x}_i, \mathbf{x}_j\right) \\ &= \int V(\mathbf{x})\, d\mu_N(\mathbf{x}) + \frac{1}{2}\int_{\neq} W(\mathbf{x}, \mathbf{y})\, d\mu_N(\mathbf{x})\, d\mu_N(\mathbf{y})\end{aligned}$$

where μ_N is the empirical measure of the particles (global encoding of the particle system)

$$\mu_N := \frac{1}{N}\sum_{k=1}^{N} \delta_{\mathbf{x}_k}$$

The model is mean field in the sense that each particle interacts with the others only via the empirical measure of the system. If $1 \leq d \leq 2$, then one can construct a random normal matrix which admits our particles at $\mathbf{x}_1, \ldots, \mathbf{x}_N$ as eigenvalues: for any $n \times n$ unitary matrix $\mathbf{U}$,

$$\mathbf{M} = \mathbf{U}\,\mathrm{diag}\left(\mathbf{x}_1, \ldots, \mathbf{x}_N\right)\mathbf{U}^H$$

which is unitary invariant if $\mathbf{U}$ is Haar distributed. Here we are more interested in an arbitrarily high dimension d, for which no matrix model is available. We make our particles at $\mathbf{x}_1, \ldots, \mathbf{x}_N$, random by considering the exchangeable probability measure P_N on $\left(\mathbb{R}^d\right)^N$ with density proportional to

$$\exp\left(-\beta_N \mathcal{I}_N\left(\mathbf{x}_1, \ldots, \mathbf{x}_N\right)\right)$$

where $\beta_N > 0$ is a positive parameter that may depend on N. The law P_N is a Boltzmann measure at inverse temperature β_N, and takes the form $\prod_{i=1}^{N} f_1\left(\mathbf{x}_i\right) \prod_{1\leqslant i<j\leqslant N} f_2\left(\mathbf{x}_i, \mathbf{x}_j\right)$ due to the structure and symmetries of $\mathcal{I}_N$.

The model contains the complex Ginibre ensemble of random matrices as the special case

$$d = 2, \beta_N = N^2, V(\mathbf{x}) = |\mathbf{x}|^2, W(\mathbf{x}, \mathbf{y}) = 2\log\frac{1}{|\mathbf{x} - \mathbf{y}|}$$

which is two dimensional, with quadratic confinement, Coulomb repulsion, and temperature $1/N^2$. Here we denote by $|\cdot|$ the Euclidean norm of $\mathbb{R}^d$.

Beyond this two-dimensional example, the typical interaction potential W that we may consider is the Coulomb interaction in arbitrary dimension

$$W(\mathbf{x},\mathbf{y}) = K_{\Delta}(\mathbf{x}-\mathbf{y}) \text{ with } K_{\Delta}(\mathbf{x}) = \begin{cases} |\mathbf{x}| & ifd = 1 \\ \log\frac{1}{|\mathbf{x}|} & ifd = 2 \\ \frac{1}{|\mathbf{x}|^{d-2}} & ifd \geqslant 3 \end{cases}$$

and the Riesz interaction, $0 < \alpha < d$ (Coulomb if $d \geq 3$ and $\alpha = 2$) $d \geq 1$

$$W(\mathbf{x},\mathbf{y}) = K_{\Delta_\alpha}(\mathbf{x}-\mathbf{y}) \text{ with } K_{\Delta_\alpha}(\mathbf{x}) = \frac{1}{|\mathbf{x}|^{d-\alpha}}$$

The Coulomb kernel K_{Δ} is the fundamental solution of the Laplace equation, whereas the Riesz kernel K_{Δ_α} is the fundamental solution of the fractional Laplace equation, hence the notations. In other words, in the sense of Schwartz–Sobolev distributions, for some constant c_d,

$$\Delta_\alpha K_{\Delta_\alpha} = c_d \delta_0$$

If $\alpha \neq 2$, then the operator Δ_α is a nonlocal Fourier multiplier.

1.4 A Mathematical Theory of Big Data

This section presents a mathematical theory to unify big data systems. Basic questions for big data includes:

- What is the theoretical foundation of big data?
- The science of data or the science of information?
- What is information?
- Are the definitions of information given by Shannon and Von Neumann sufficient for big data?
- How is "free entropy" relevant to the new definition of "information"?

Applications of big data include: (i) quantum systems; (ii) financial systems; (iii) atmospheric systems; (iv) sensor network (e.g., PMU, WAMS); (v) wireless networks (vehicle-to-vehicle communications, 5G); (vi) transportation; (vii) manufacturing; (viii) health (patients), and so forth.

The big picture of research is the interaction of random matrices, geometric functional analysis, and algorithms (theoretical computer science). We make the following observations:

- Random matrices are natural building blocks to model big data.
- At the heart of random matrix theory lies the realization that the *spectrum* of a random matrix **X** tends to stabilize as the dimensions of **X** grows to infinity.
- In the last few years, considerable progress was made on the more difficult local and nonasymptotic regimes. In the nonasymptotic regimes, the dimensions of **X** are **fixed** rather than grow to infinity.
- Connections among random matrix theory, quantum information theory, free probability, and statistics complete the picture.

The central objective of this section is to establish the fact that the circular law is the consequence of the more basic concept of "free entropy". Here we only sketch the key conceptual steps that complete the proof [1].

Circular and ring laws for eigenvalues are fundamental to random matrices. Non-Hermitian random matrices, and thus their eigenvalues, are complex values. See Chapter 6 for details. The circular law is observed for the (square) complex i.i.d. ensemble, while the ring law is for the rectangular complex i.i.d. ensemble. For an $N \times T$ complex matrix, the inner radius is $\sqrt{1-c}$, where $c = N/T \leq 1$. The circular law is the special case of the rectangular law for $N = T$ or $c = 1$.

The circular law [56] states that the empirical measure of the eigenvalues of a random $n \times n$ matrix, with i.i.d. entries of variance $1/n$, tends to the uniform law on the unit disc of the complex plane, as the dimension n tends to infinity. This universal result was proved rigorously by Tao and Vu [55], after 50 years of contributions. The circular law is universal, in the sense that it remains valid if one drops the Gaussian assumption of the entries of the matrix, while keeping the i.i.d. structure and the $1/n$ variance. The proof of this high dimensional phenomenon involves tools from potential theory, from additive combinatorics, and from asymptotic geometric analysis. The circular law phenomenon can be checked in the Gaussian case using the fact that the model is then exactly solvable. Actually, Ginibre has shown in the 1960s that if the entries are i.i.d.-centered complex Gaussians then the eigenvalues of such matrices form a Coulomb gas at temperature $1/n$ in dimension 2. This in turn suggests exploration of the analog of the circular law phenomenon in dimension ≥ 3, beyond random matrices. This led researchers to introduce in [57] stochastic interacting particle systems in which each particle is confined by an external field, and each pair of particles is subject to a singular repulsion. Under general assumptions and suitable scaling, the empirical measure of the particles converges, as the number of particles tends to infinity, to a probability measure that minimizes a natural energy-entropy functional. In the case of quadratic confinement and Coulomb repulsion, the limiting law is uniform on a ball.

Non-Hermitian matrices have a complex-valued eigenvalue distribution in general. In the Hermitian case, we work on the complex-valued matrix functions to search for real-valued eigenvalues, while we now have to work on a q-valued function to search complex-valued eigenvalues. See Table 1.1. For large non-Hermitian random matrices, we need quatartenionic free probability theory.

Boltzmann entropy (statistical physics), Shannon entropy (for classical information theory) and von Neumann entropy [39] (for quantum information) are all defined on the set of positive real-valued numbers. The eigenvalues of non-Hermitian random matrices are complex valued, in general. This basic fact suggests that the concepts and hence formulations based on Boltzmann entropy, Shannon entropy, and von Neumann entropy may not be sufficient for the theory of big data based on non-Hermitian random matrices.

Von Neumann entropy is defined as [58]

$$S(\rho) = \mathrm{Tr}\, \phi(\rho) = -\sum_{i=1}^{n} \lambda_i \log \lambda_i$$

1 The similar justification of the ring law from a more basic concept is open at this point of writing.

Table 1.1 Comparison between classical, free, and quatartenionic free probability theories.

	Probability space	Algebra
Classical probability	Commutative	Commutative
Free probability	Noncommutative	Commutative
Quatartenionic free probability	Noncommutative	Noncommutative

Table 1.2 Comparison of different entropy definitions.

	Definition set	Mathematical expression	Remarks
Shannon/Boltzmann entropy	Positive real values	$S(\mathbf{p}) = -\sum_{i=1}^{n} p_i \log p_i.$	p_i are positive
Von Neumann entropy	Positive real values	$S(\rho) = \mathrm{Tr}\, \phi(\rho) = -\sum_{i=1}^{n} \lambda_i \log \lambda_i.$	λ_i are positive
Free entropy	Complex values	$\chi(\mu) := \iint \log \lvert x-y \rvert \mu(dx)\mu(dy).$	μ is complex on $\mathbb{C}$

where λ_i are the eigenvalues of ρ, a statistical operator, and $\phi : \mathbb{R}^+ \to \mathbb{R}$ is the continuous function $\phi(t) = -t \log t$. When studying non-Hermitian random matrices, we find that the eigenvalues λ_i are complex values, instead of real (positive) values. This suggests that von Neumann entropy is insufficient for the non-Hermitian data matrices. It is well known that Shannon entropy may be viewed as a special case of von Neumann entropy.

1.4.1 Boltzmann Entropy and H-Theorem

Consider a system of n distinguishable particles, each of them being in one of r possible states (typically energy levels). We have $n = n_1 + \cdots + n_r$ where n_i is the number of particles in state i. $S(\mathbf{p}) = -\sum_{i=1}^{n} p_i \log p_i$, where $\mathbf{p} := (p_1, \ldots, p_r)$. The quantity $S(p)$ is the Boltzmann entropy of the discrete probability distribution p. It appears here as an asymptotic additive degree of freedom per particle in a system with an infinite number of particles, each of them being in one of the r possible states, with population frequencies $p_1, ..., p_r$.

Returning to the motivations of Boltzmann, let us recall that the first principle of Carnot–Clausius thermodynamics states that the internal energy of an isolated system is constant, and the second principle states that there exists an extensive state variable called the entropy that can never decrease for an isolated system. Boltzmann wanted to derive the second principle from the idea (controversial, at that time) that matter is made with atoms. The H-theorem states that the entropy $S = -H$ is *monotonic* along the Boltzmann equation.

1.4.2 Shannon Entropy and Classical Information Theory

Boltzmann entropy also plays a fundamental role in communication theory [59]. It was founded in the 1940s by Claude Elwood Shannon (1916–2001) at Bell Labs, where it is known as "Shannon entropy."

> My greatest concern was what to call it. I thought of calling it "information", but the word was overly used, so I decided to call it "uncertainty". When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, "You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage." (Claude E. Shannon, 1961)

1.4.3 Dan-Virgil Voiculescu and Free Central Limit Theorem

Free probability theory was forged in the 1980s by Dan-Virgil Voiculescu (1946–), while working on isomorphism problems in von Neumann operator algebras of free groups. Voiculescu discovered in the 1990s that free probability is the algebraic structure that appears naturally in the asymptotic global spectral analysis of random-matrix models as the dimension tends to infinity. Free probability theory comes with algebraic analogs of the central limit theorem and the Boltzmann entropy.

For the $n \times n$ complex matrix $\mathbf{A} \in \mathcal{M}_n(\mathbb{C})$, τ appears as an expectation with respect to the empirical spectral distribution. Denoting $\lambda_1(\mathbf{A}), \ldots, \lambda_n(\mathbf{A}) \in \mathbb{C}$ the eigenvalues of $\mathbf{A}$, we have

$$\tau(\mathbf{A}) = \frac{1}{n}\sum_{k=1}^{n} \delta_{\lambda_k(\mathbf{A})} = \int x \mu_{\mathbf{A}}(dx), \text{ where } \mu_{\mathbf{A}} := \frac{1}{n}\sum_{k=1}^{n} \delta_{\lambda_k(\mathbf{A})}$$

We also obtain

$$\begin{aligned} 2\tau\left(\log\left((\mathbf{A}-z\mathbf{I})(\mathbf{A}-z\mathbf{I})^*\right)\right) &= \frac{1}{n}\log\left|\det\left(\mathbf{A}-z\mathbf{I}_n\right)\right| \\ &= \int \log|z-\lambda|\, d\mu_{\mathbf{A}}(\lambda) \\ &= \left(\log|z-\cdot| * \mu_{\mathbf{A}}\right)(z) \\ &=: -U_{\mu_{\mathbf{A}}}(z) \end{aligned}$$

The quantity $U_{\mu_{\mathbf{A}}}(z)$ is exactly the logarithmic potential at point $z \in \mathbb{C}$ of the probability measure $\mu_{\mathbf{A}}$.

Since $-\frac{1}{2\pi}\log|z-\cdot|$ is the so-called fundamental solution of the Laplace equation in dimension 2, it follows that, in the sense of Schwartz–Sobolev distributions, $\mu_{\mathbf{A}} = \frac{1}{2\pi}\Delta U_{\mu_{\mathbf{A}}}$. It is amazing to point out that the (discrete) empirical spectral distribution follows a (continuous) partial differential equation—the Laplace equation.

1.4.4 Free Entropy

Inspired by Boltzmann and Shannon on the central limit theory (CLT) of classical probability theory, we may ask if there exists, in free probability theory, a free entropy

functional, maximized by the semicircle law at fixed second moment, and which is monotonic along the free CLT.

The semicircle law is, for the free entropy, the analog of the Gaussian law for the Boltzmann entropy. The semicircle law on $[-2, 2]$ is the unique law that maximizes the Voiculescu entropy χ among the laws on $\mathbb{R}$ with a second moment equal to 1, for $\text{supp}(\mu) \subset \mathbb{R}$,

$$\arg\max\left\{\chi(\mu) : \int x^2\mu(dx) = 1\right\} = \frac{1}{2\pi}\sqrt{4-x^2}\mathbf{1}_{[-2,2]}(x)\,dx$$

How about laws on $\mathbb{C}$ (complex values) instead of $\mathbb{R}$ (real values)? When μ is a probability measure on $\mathbb{C}$, we will denote the Voiculescu entropy functional as

$$\chi(\mu) := \iint \log|x-y|\,\mu(dx)\mu\left(dy\right)$$

The uniform law on the unit disc is the unique law that maximizes the functional χ among the set of laws on $\mathbb{C}$ with the second moment (mean squared modulus) equal to 1 (here $z = x + iy$, and $dz = dxdy$) for $\text{supp}(\mu) \subset \mathbb{C}$

$$\arg\max\left\{\chi(\mu) : \int |z|^2\mu(dz) = 1\right\} = \frac{1}{\pi}\mathbf{1}_{\{z\in\mathbb{C}:|z|=1\}}dz$$

This phenomenon is known as the circular law. Under the uniform law on the unit disc, the real and the imaginary parts follow the semicircle law on $[-1, 1]$, and are not independent.

If one starts with a Hermitian random Gaussian matrix, the Gaussian unitary ensemble (GUE), then the same analysis is available, and produces a convergence to the semicircle law on $[-2, 2]$.

It turns out that the Voiculescu free entropy χ is monotonic along the Voiculescu free CLT. The Boltzmann–Shannon H-theorem interpretation of the CLT is thus remarkably valid in classical probability theory, and in free probability theory.

$\mathbf{A}$ and $\mathbf{B}$ are two $n \times n$ Hermitian matrices such that $\mu_{\mathbf{A}} \to \mu_a$, and $\mu_{\mathbf{B}} \to \mu_b$, in the sense of moments as $n \to \infty$, where μ_a and μ_b are two compactly supported laws on $\mathbb{R}$. Let $\mathbf{U}$ and $\mathbf{V}$ be independent random unitary matrices uniformly distributed on the unitary group (we say Haar unitary). Then

$$\mathbb{E}\mu_{\mathbf{UAU}^*+\mathbf{VBV}^*} \xrightarrow[n\to\infty]{*} \mu_a \boxplus \mu_b$$

This asymptotic freeness reveals that free probability is the algebraic structure that emerges from asymptotic analysis of large dimensional unitary invariant models of random matrices. As the functional χ is maximized by the uniform law on the unit disc, one may ask about an analog of the Wigner theorem for non-Hermitian random matrices. The answer is positive.

1.4.5 Jean Ginibre and his Ensemble of Non-Hermitian Random Matrices

The circular law for the Complex Ginibre ensemble, can be proved using the Voiculescu functional χ (maximized at fixed second moment by uniform law on unit disc). A simple model of random matrix is the Ginibre model:

$$\mathbf{G} = \begin{pmatrix} G_{11} & \cdots & G_{1n} \\ \vdots & \vdots & \vdots \\ G_{n1} & \cdots & G_{nn} \end{pmatrix}$$

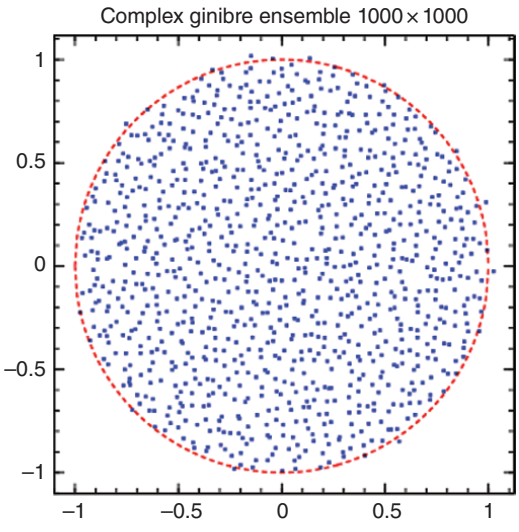


Figure 1.4 The eigenvalues of a single matrix drawn from the complex Ginibre ensemble of random matrices. The dashed line is the unit circle. This numerical experiment was performed using the promising Julia http://julialang.org/ (accessed August 17, 2016).

where $\left(G_{jk}\right)_{1\leqslant j,k\leqslant n}$ are i.i.d. random variables on $\mathbb{C}$, with $\operatorname{Re} G_{jk}, \operatorname{Im} G_{jk}$ of the Gaussian law of mean 0 and variance $1/(2n)$. The eigenvalues of a single matrix drawn from the complex Ginibre ensemble of random matrices is illustrated in Figure 1.4.

The density of $\mathbf{G}$ is proportional to

$$\prod_{j,k=1}^{n} \exp\left(-n\left|G_{jk}\right|^2\right) = \exp\left(-\sum_{j,k=1}^{n} n\left|G_{jk}\right|^2\right) = \exp\left(-n\operatorname{Tr}\left(\mathbf{G}\mathbf{G}^H\right)\right)$$

1.4.6 Circular Law for the Complex Ginibre Ensemble

The law of the eigenvalues is then proportional to

$$\exp\left(-n\sum_{j=1}^{n}\left|\lambda_j\right|^2\right)\prod_{1\leqslant j,k\leqslant n}\left|\lambda_j-\lambda_k\right|^2$$

This defines a determinantal process on $\mathbb{C}$: the complex Ginibre ensemble. In order to interpret the law of the eigenvalues as a Boltzmann measure, we put the Vandermonde determinant inside the exponential:

$$\exp\left(-n\sum_{j=1}^{n}\left|\lambda_j\right|^2+2\sum_{j<k}\log\left|\lambda_j-\lambda_k\right|\right)$$

If we encode the eigenvalues by the empirical measure

$$\mu_n := \frac{1}{n}\sum_{j=1}^{n}\delta_{\lambda_j}$$

this takes the form

$$e^{-n^2 I(\mu_n)}$$

where the "energy" $\mathcal{I}\left(\mu_n\right)$ of the configuration μ_n is defined via

$$\mathcal{I}\left(\mu_n\right) := \int |z|^2 d\mu\left(z\right) + \iint_{\neq} \log \frac{1}{|z - z'|} d\mu\left(z\right) d\mu\left(z'\right)$$

This suggests interpreting the eigenvalues $\lambda_1, \ldots, \lambda_n$ of $\mathbf{G}$ as Coulomb gas of two-dimensional charged particles, confined by an an external field (quadratic potential) and subject to pair Coulomb repulsion.

$-\mathcal{I}$ can also be seen as a penalized Voiculescu functional. Minimizing a penalized functional is equivalent to minimizing without penalty but under constraint (Lagrange). Presently, if $\mathcal{M}$ is the set of probability measures on $\mathbb{C}$ then $\inf_{\mathcal{M}} \mathcal{I} > -\infty$ and the infimum is achieved at a *unique* probability measure μ_*, which is the **uniform** law on the unit disc of $\mathbb{C}$. The circular law is *universal*, in the sense that it remains valid if one drops the Gaussian assumption of the entries of the matrix, while keeping the i.i.d. structure and the $1/n$ variance.

How does the random discrete probability measure μ_n behave as $n \to \infty$? We may adopt a large deviations approach. Let $\mathcal{M}$ be the set of probability measures on $\mathbb{C}$. We may show that the functional $\mathcal{I} : \mathcal{M} \to \mathbb{R} \cup \{+\infty\}$ is lower semicontinuous for the topology of narrow convergence, is strictly convex, and has compact level sets. Let us consider a distance compatible with the topology. It can be shown that for every ball B for this distance

$$\mathbb{P}\left(\mu_n \in B\right) \approx \exp\left(-n^2 \inf_B \left(\mathcal{I} - \inf_B \mathcal{I}\right)\right)$$

The first Borel–Cantelli lemma allows one to deduce that almost surely

$$\lim_{n \to 0} \mu_n = \mu_* = \arg\inf \mathcal{I} = \frac{1}{\pi} \mathbf{1}_{\{z \in \mathbb{C} : |z| \leqslant 1\}} dz$$

where $z = x + iy$ and $dz = dxdy$. This phenomenon is known as the circular law. If one starts with a Hermitian random Gaussian matrix—the Gaussian unitary ensemble (GUE)—then the same analysis is available, and produces a convergence to the semicircle law on $[-2, 2]$.

1.5 Smart Grid

Roughly speaking, a smart grid can be viewed as two flows: (i) information, and (ii) electric power. The information flow is used for grid control. Communications, sensing, and control must be considered jointly. At an abstract level, the smart grid can be viewed as an "energy Internet." This is very relevant to the Internet of Things, for machine-to-machine communications.

The vision of a smart transmission grid is illustrated in Figure 1.5. As a roadmap for research and development, the smart features of the transmission grid are envisaged and summarized as digitization, flexibility, intelligence, resilience, sustainability, and customization. The enabling technologies include [60]:

- *New materials and alternative clean energy resources.* The high penetration of alternative clean energy resources will mitigate the conflicts between the development of human society and environment sustainability.

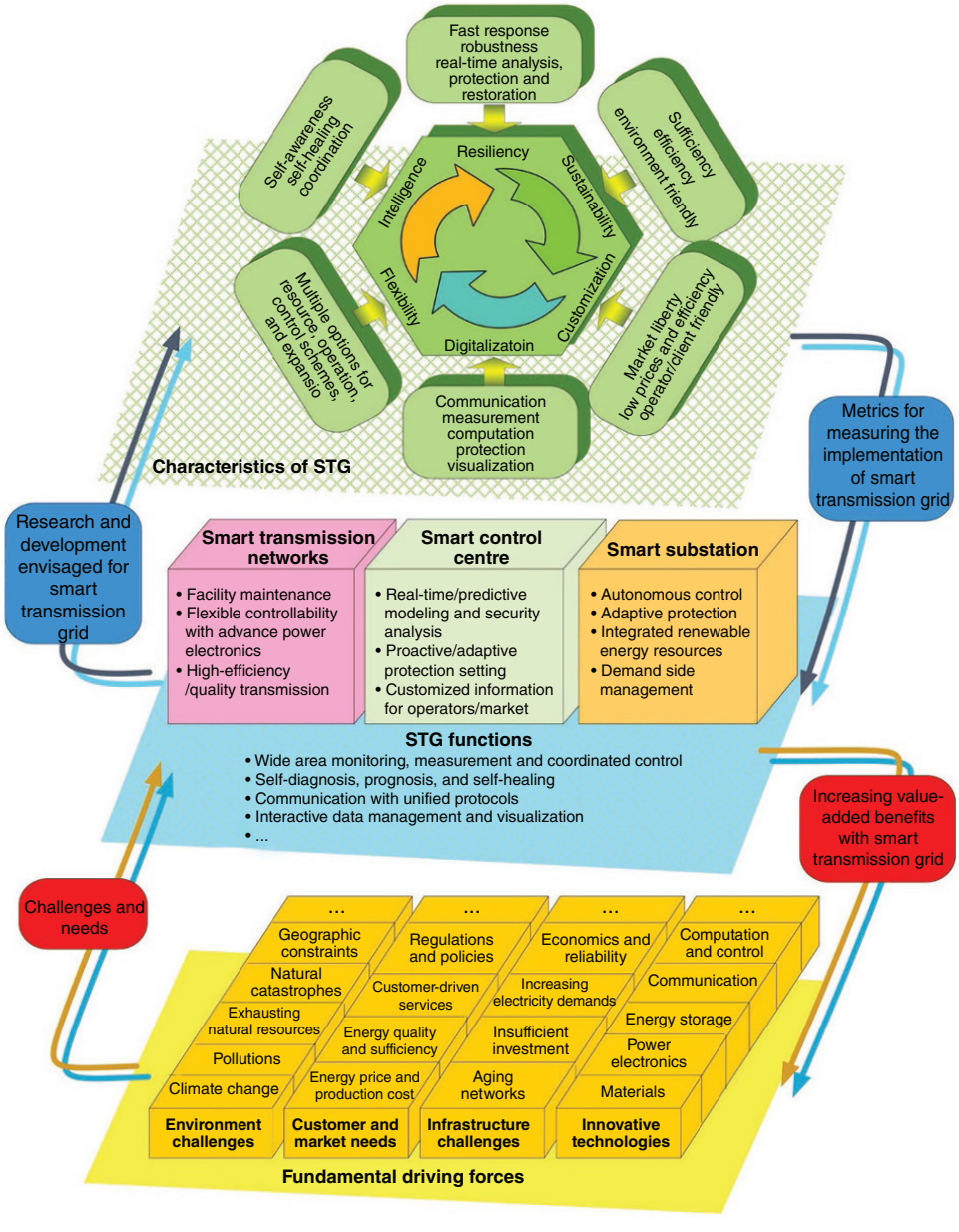


Figure 1.5 Vision of a smart transmission grid. Source: Reproduced from [60] with Permission of IEEE.

- *Advanced power electronics and devices.* These greatly improve the quality of power supply and flexibility of power flow control.
- *Sensing and measurement.* The basis for communications, computing, control, and intelligence.
- *Communications*. Adaptive communication networks will allow open-standardized communication protocols to operate on a unique platform. Real-time control based on fast and accurate information exchange on different platforms will improve system resilience by the enhancement of system reliability and security, and optimization of the transmission asset utilization.
- *Advanced computing and control methodologies*. High-performance computing, parallel, and distributed computing technologies will enable real-time modeling and simulation of complex power systems. The accuracy of the situation awareness will be improved for further suitable operations and control strategies. Advanced control methodologies and novel distributed control paradigms will be needed to automate the entire customer-centric power-delivery network.
- *Mature power market regulation and policies*. These improve the transparency, liberty, and competition of the power market. High customer interaction with the electricity consumption should be enabled and encouraged.
- *Intelligent technologies.* These enable fuzzy logic reasoning, and knowledge discovery.

1.6 Big Data and Smart Grid

Our knowledge is dominated by the scales in which our observations are made. Our slogan is "data is science and science is data." This book treats big data as the foundation for the smart grid, an approach that is consistent with [39] and [40]. In other words, the science of smart grid is a combination of distributed sensing and a distributed network with the electric power grid. See Chapter 11 for details about why big data should be tied together with the smart grid. The central task is to understand the statistical knowledge of the massive datasets and make sense of these data.

Large random matrices are used to model large datasets. It is our firm belief that large random matrices are the basic building blocks for our science. It is the calculus for data. From the point of view of probability and statistics, after living in the age of vector-valued random variables, we are entering a new age of big data, an age of matrix-valued random variables. Initially, Newton and Leibniz developed the calculus of $f(x)$, where x is a free variable. Later, we study $f(X)$ where is a scalar-valued random variable (a function defined on the sample space). Then, we study $f(\mathbf{x})$ where $\mathbf{x} = \left[X_1, \ldots, X_N\right]^T$ is a vector-valued random variable. Now we are entering an age of studying $f(\mathbf{X})$ where $\mathbf{X} = \left(\mathbf{x}_1, \ldots, \mathbf{x}_n\right) \in \mathbb{C}^{N\times n}$ is a matrix-valued random variable. In particular, we are interested in the asymptotic regime of

$$N \to \infty, n \to \infty \text{ but } \frac{N}{n} \to c \in (0, \infty)$$

Alternatively, we are interested in the nonasymptotic regime where

N, n are large but finite

For a complex quantum system (a system with many degrees of freedom)—such as atoms, nuclei, fundamental particles, it is almost impossible to imagine a theory that

is exploitable enough to compute accurately, for instance, the energy levels of such a system. Antenna sensors, smart meters, PMUs, and stocks are analogies. The model of random particles [56] is relevant in this context, for example. Energy and entropy are two drivers.

1.7 Reading Guide

The core material of this book provides a comprehensive study of large random matrices for big data applications (Part I) (see Figure 1.6). After this has been accomplished, we make connections with selected smart grid applications (Part II) and selected applications in communications and sensing (Part III). Random matrix theory has been used as the unifying tool to tie the three parts together. Very often, connections are made at the mathematical level.

Missing links are, however, inevitably frequent because the majority of materials (90% we guess) appear, for the first time, in book form. Even worse, most materials are treated, for the first time, in the context of engineering applications. The main obstacle when reading this book is the mathematical depth. Although tested in the class room, the limited size of this book makes it impossible to present all the material in a self-contained manner.

For the large random matrix, the trick is to convert two-dimensional matrix problems into one-dimensional problems by using the eigenvalue distribution—forget about the eigenvectors for the moment. As a result, we can study the function of $f\left(\lambda_i\right), i = 1, \ldots, n$, for some function f. Various functions f are defined for different applications.

To interpret our empirical discovery in [61] we connected our results with quantum information theory; see [39] for this link. About 200 pages were also dedicated to random matrix theory in [39]. It was realized that our discovery was caused by the high dimensionality of the problem, which lead to the "concentration of measure" phenomenon, a high-dimensional effect, or a property of a large number of variables, for which functions with small local oscillations are almost constant. In this connection,

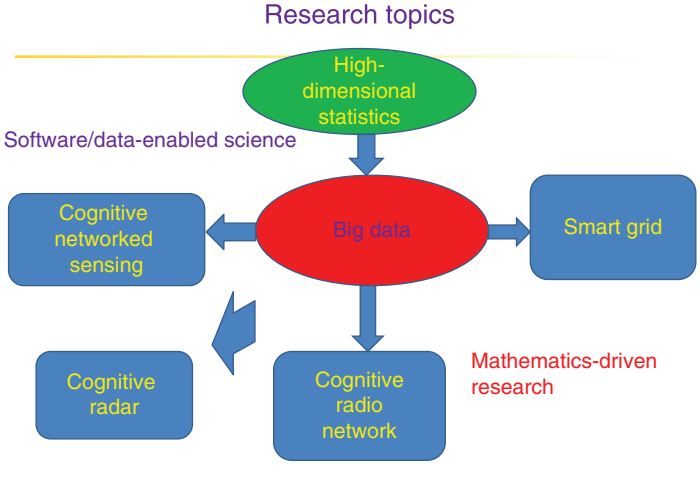


Figure 1.6 Big data vision.

the first author's book [40] was born. The current book can be viewed applying [39] and [40]. The use of random matrix theory as the unifying theme to model large wireless networks, smart grid and big data was explicitly pointed out in [39]. [40] was written to support this big vision. These three books are complementary. We closed the circle during the writing of three books. Now we are revisiting random matrix theory, with the emphasis on the latest results, which are also applicable to the problems we have in mind—smart grid and big data. During this adventure, the most remarkable experience with random matrix theory is our feeling of being shocked by its usefulness, beauty, depth and fertility, as pointed out in the preface of [62]. According to him, usefulness is usually measured by the utility of the topic outside mathematics. Beauty is a quality of much the material, but is often something only a trained eye can see. "Depth comes via the linking together of multiple ideas and topics, often seemingly removed from the original context. And fertility means that with a reasonable effort there are new results, some useful, some with beauty, and a few maybe with depth, still waiting to be found."

In Chapter 1, we started our book with some challenges for big data. Chapter 2 gives an overview of the framework for the mathematical framework needed for the analysis of big data. We use a bottom-up approach to lay the foundations using large random matrices to summarize the large datasets (big data).

Chapter 3 gives the fundamentals of large dimensional random matrices. One motivation is to model the large datasets using large random matrices. It is our belief that large dimensional random matrices are the foundation for the analysis of big data; this chapter is the basic material for next-generation engineers and researchers.

Chapter 4, by studying the central limit theory for linear spectral statistics, addresses the spectral analysis of large dimensional random matrices. The main reason is because many important statistics in multivariate statistical analysis can be expressed as functionals of the empirical spectral distribution of some random matrices.

Chapter 5 studies the Hermtian free probability theory. The idea of exploiting "large models" is the unified theme of this whole book. As a result, large random matrices are natural building blocks for the entire theoretical framework. As large random matrices can be regarded as free random variables, matrix-valued free probability is discussed to study the variables.

Chapter 6 studies (large) non-Hermitian random matrices using the newly developed quatartenionic free probability theory. Most results appear in book form for the first time.

Chapter 7 deals with data collection. Data storage is central to big data. For many applications, we often cannot afford the luxury of saving all the raw data generated by the system (or network) for future processing. One fundamental challenge is to choose which types of information are stored. As we deal with streaming data, real-time processing is required.

Chapter 8 deals with anomaly detection using large random matrices. One objective is is to study the denial of service using big data. We understand how the large data size affects the matrix hypothesis detection.

Requirements for applying big data to smart grid are addressed in Chapter 9. The technical challenges are discussed in Chapter 10. And big data topics for smart grid are addressed in Chapter 11.

Chapter 12 introduces grid monitoring and state estimation using phasor measurement units (PMUs). Chapter 13 gives an exhaustive treatment of false data injection

attacks in the context of state estimation. It is well known that cyber security is the most important task facing engineers and researchers. We use false data injection to attack against state estimation.

Chapter 14 briefly discusses the demand response.

Chapter 15 addresses communications topics for smart grids. To control the power grid we need sensing and communications to tie together the whole grid. High-performance computing and distributed computing are two enablers.

Bibliographical Remarks

This current book together with another two books [39,40] pursues a paradigm of modeling big data using large random matrices. To the best of our knowledge, this vision was explicitly spelled out and formulated analytically for the first time in November of 2011 during the writing of [39].

Section 1.1.5 draws on material from [4, 26, 27]. We are now facing the data deluge. [63]. For big data we follow [22], which is an excellent review and tutorial. Labrinidis and Jagadish (2012) [24] is very insightful. We have followed [24] for insights.

The state-of-the-art of big data is that there is no clear definition for big data, or the adopted theoretical framework. Our aim of these three books is to attempt to define our big data problems in a random matrix way. There is no claim of solving all big data problems using one framework. In Section 1.3 (Definition 1.3.1), we use three conditions to define our problems related to big data. We limit the potential applications of our methods using Definition 1.3.1. Clarity and rigor, on the other hand, are achieved. Our three books are aimed at addressing the consequences of Definition 1.3.1 in the context of large random matrices.

We have drawn from [45] for some parts of Section 1.2. For Example 1.2.3, we also drew from [47]. Challenges include: (i) Real-time processing [64] is challenging; (ii) other technical challenges [65]; (iii) signal processing [66].

Example 1.3.2 and Example 1.3.3 are adapted from [49, 67].

We adopt statistical methods to study big data. Fisher [68] states that the purpose of statistical methods is to reduce a large quantity of data to a small amount of data that is capable of containing as much of the relevant information as possible in the original data. Because the data will generally supply a large number of "facts," many more than are sought, much information in the data is irrelevant. This brings to the fore the Fisherial dictum [69, p. 1] that statistical analysis via the reduction of the data is the process of extracting the irrelevant information. This may be accomplished by modeling a hypothetical population specified by relative few parameters. See [44,70] for one application in modeling big data in large wireless networks.

Functions are the core for the practical applications. Tao's excellent text [67] relies heavily on Talagrand's concentration theorem for convex functions. High-dimensional spaces were used to model big data, originally in [39] and then in [40], by using large random matrices.

In Example 1.3.7, we draw some material from [71]. Example 1.3.8 takes results from [72–75]. More recent work is [76–78]. In [44,70] we used non-Hermitian random matrices to model big data collected in a large-scale wireless cognitive radio network. We follow [54] in Example 1.3.9.

Example 1.3.10 follows [79].

Example 1.3.11 follows [80].

Example 1.3.12 is taken from [81].

Example 1.3.16 follows [56, 57].

In Section 1.4 we follow [56] for the development of the unified mathematical theory for big data.

Part I

Fundamentals of Big Data

2

The Mathematical Foundations of Big Data Systems

This chapter gives an overview of the mathematical framework needed for the analysis of big data. Some topics are only listed. Some chapters covered later in this book aim to go deeper in some selected directions that will be relevant to both power grids and big data. In particular, we use a bottom-up approach to lay the foundations using large random matrices to summarize the large datasets (big data). Random matrices play a central role to describe a null hypothesis or a *minimum information hypothesis* for the description of a large big data system or subsystem.

In big data we deal with new signal and information processing methods that can capture, analyze, and represent emerging datasets that do not fall into traditional "signal" categories (such as speech or video). Once we store, index and query very large datasets using parallel and distributed computing systems, we then mine and extract knowledge from these very large datasets to obtain big data analytics.

Random matrix models provide a powerful framework for modeling numerous physical phenomena, with applications covering all branches of theoretical physics. Finding correlations between observables is at the heart of scientific methodology. Once correlations between "causes" and "effects" are *empirically* established, one can start devising theoretical models to understand the mechanisms underlying such correlations, and use these models for prediction purposes. In many cases, for example in financial systems [82], the number of possible causes and resulting effects are large. In financial systems, it is suggested that "large models" should be at the *forefront* of the econometrics agenda. We adopt this viewpoint in big data systems.

The Marchenko–Pastur law is naturally used to model the large data sets represented in terms of large random matrices. The natural generalizations of the Marchenko–Pastur law include free random variables and data with power-law tails. These explicit expressions—benchmarks—are valid for the interval where singular values are expected in the *absence of any true correlations* (or null hypothesis $\mathbb{H}_0$) between the variables under study. Any deviation from these benchmarks indicates "signal"—the *presence of any true correlations* (the alternative signal hypotheses $\mathbb{H}_1$). Our central goal for big data is to distinguish "signal" from "noise." As a result, understanding these benchmarks is central to this book. The systems under study are complex; the classical assumption of linearity of the systems is, in general, invalid.

After introducing some basics, we study, as examples, a number of big data systems: (i) quantum system; (ii) financial system; (iii) atmospheric system; (iv) sensing network; (v) wireless network; (vi) smart grid; (vii) transportation. Historically, the quantum system and the financial system have been most widely studied. These systems are

Smart Grid using Big Data Analytics: A Random Matrix Theory Approach, First Edition.
Robert C. Qiu and Paul Antonik.

unified through the use of large random matrices. Mathematically speaking, we deal with matrix-valued random variables.

2.1 Big Data Analytics

Curiously enough, big data was a serious problem just a few years ago. When data volumes started skyrocketing in the early 2000s, storage and CPU technologies were overwhelmed by the numerous terabytes of big data—to the point that traditional signal and information processing methods were invalid. Storage and CPUs not only developed greater capacity, speed, and intelligence; they also fell in price. Enterprises went from being unable to afford or manage big data to lavishing budgets on its collection and analysis.

Today, many engineering projects are exploring big data to discover facts that were unknown before. Using advanced analytics, industry can study big data to understand the current state of the business and track still-evolving aspects such as customer behavior.

Big data analytics is where advanced analytic techniques operate on big data sets. Hence, big data analytics is really about two things—big data and analytics. The real power comes from the combination of both things. Big data analytics is the application of advanced analytic techniques to very big data sets. First, there is big data for massive amounts of detailed information. Second, there is advanced analytics, which is actually a collection of different tool types, including those based on predictive analytics, data mining, statistics, artificial intelligence, natural language processing, and so on. Put them together and you get big data analytics, the hottest new practice in industry.

In this book, our viewpoint is built upon the mathematical objects of large random matrices and we use them for a unified framework of analysis. Our aim is to promote this unified framework using large random matrices, rather than a collection of different tool types. First the large datasets are captured (and stored for easy indexing and querying), then they are represented by the notion of large random matrices with a minimum information hypothesis of these datasets. Finally, analytics are obtained from these large random matrices, such as the the limiting distribution of eigenvalues (when the sizes of these random matrices approach infinity). We deal with matrix-valued functions (or analytics) for a time-indexed sequence of large random matrices $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ of size $N \times T$ for time $t = T, \ldots, nT$. often $N = 100 - 10000$ and $T = 100 - 10000$. Here N random variables are considered jointly. One primary advantage of using random matrix theory is its *universality* in the sense that the results are valid for arbitrary distributions of these N random variables. In practice, this is critical because we often do not know the distributions of these random variables—the data is messy.

The definition is easy to understand but do users actually use the term? To quantify this question, the survey for the report [83] asked: "Which of the following best characterizes your familiarity with big data analytics and how you name it?" Among 325 respondents, 65% said "I know what you mean, but I do not have a formal name for it," and 28% said "I know what you mean, and I have a name for it." Only 7% say "I have not seen or heard of anything resembling big data analytics." When users have a term, it is most often "big data analytics."

Why put big data and analytics together now?

- Big data provides gigantic statistical samples, which enhance analytic tool results.
- Analytic tools and databases can now handle big data.
- The economics of analytics is now more embraceable than ever.
- There is a lot to learn from messy data, as long as it is big. Discovery and predictive analytics depend on lots of details—even questionable data. Data are often missing.
- Big data is a special asset that merits leverage.
- Analytics based on large data samples reveals and leverages business change.

In our context, big data are represented by random matrices $\mathbf{X}_i$, $i = 1, \ldots, n$ and the analytics are matrix-valued functions $f\left(\mathbf{X}_1, \ldots, \mathbf{X}_n\right)$. For example, the following basic functions are natural:

- adding up the n matrices $\mathbf{A}_n = \mathbf{X}_1 + \mathbf{X} + \cdots + \mathbf{X}_n$;
- products of n matrices $\mathbf{P}_n = \mathbf{X}_1\mathbf{X}_2 \cdots \mathbf{X}_n$;
- geometric mean of n matrices $\left(\mathbf{P}_n\right)^{1/n} = \left(\mathbf{X}_1\mathbf{X}_2 \cdots \mathbf{X}_n\right)^{1/n}$;
- $\mathbf{X}_1^{1/M}\mathbf{X}_2^{1/M} \cdots \mathbf{X}_n^{1/M}$ for non-negative integer $M \geq 1$.

In general these matrices have no symmetry; they are non-Hermitian and complex. Sometimes the only knowledge we know about these matrices is to which class these matrices belong, such as independent identically distributed (i.i.d.) Gaussian. See Chapter 6 for details.

2.2 Big Data: Sense, Collect, Store, and Analyze

Mathematics for Analysis of Petascale Data, funded by DOE, addresses the mathematical challenges of extracting insight from huge scientific datasets, finding key features and understanding the relationships between those features. Research areas include machine learning, real-time analysis of streaming data, stochastic nonlinear data-reduction techniques and scalable statistical analysis techniques applicable to a broad range of DOE applications including sensor data from the electric grid, cosmology, and climate data.

Advances in digital sensors, communications, computation, and storage have created huge collections of data, capturing information of value to business, science, government, and society. In Chapter 7, we formulate the problem in terms of covariance matrix estimation, which is ultimately reduced to a problem of convex optimization. Real-time analytics and large-scale optimization parameters are challenges.

Random matrix theory fits into the framework of stochastic *nonlinear* data-reduction techniques and *scalable* statistical analysis.

Sensors: We are interested in sensing the grid using smart meters and PMUs. Communication infrastructure also generates a lot of data, for example through spectrum sensing.

Computer networks: Data from the many different sources can be collected into massive data sets via localized sensor networks, as well as the Internet. Distributed computing is often required for real-time applications.

Data storage: Advances in magnetic-disk technology have dramatically decreased the cost of storing data. For example, a 1 terabyte disk drive, holding one trillion bytes of data, costs around $100.

Cluster computer systems: A new form of computer system, consisting of thousands of "nodes," each having several processors and disks, connected by high-speed local-area networks, has become the chosen hardware configuration for data-intensive computing systems. These clusters provide both the storage capacity for large data sets, and the computing power to organize the data, to analyze it, and to respond to queries about the data from remote users. Cluster computers are designed to manage and analyze very large data sets. The "trick" is in the software algorithms.

Cloud computing facilities: The rise of large data centers and cluster computers has created a new business model, where businesses and individuals can rent storage and computing capacity, for example Amazon Web Services.

Data analysis algorithms: The enormous volumes of data require automated or semiautomated analysis—techniques to detect patterns, identify anomalies, and extract knowledge. Again, the key is in the software algorithms—new forms of computation, combining statistical analysis, optimization, and artificial intelligence to construct statistical models from large collections of data and to infer how the system should respond to new data. For example, Netflix uses machine learning in its recommendation system.

- How do we take advantage of cloud computing to instantiate big data services in an optimal manner (i.e., to reduce cost, maximize performance)?
- How do we automate and formalize the process of instantiating the entire data analysis pipeline?
- How do we track provenance and handle security as the data flows through the analysis pipeline?
- What additional storage and analysis systems do we need? For example, do we need a Hadoop for graphs? What is the role of in-memory systems?

2.2.1 Data Collection

All data that is to be processed is consolidated for analysis. Difficulties with data collection lie in the different forms that data may have as they arrive from different sources. Data integration is later performed to keep data as cohesive as possible. Data collection can be designed to facilitate the data integration.

The sheer size of the data is a challenge and also an opportunity. Cloud computing provides a solution that meets some scalability needs. The major problem with this system would be getting the data into the cloud to begin processing. Using standard Internet connections to upload the data to the cloud would be a significant bottleneck in the process.

2.2.2 Data Cleansing

After collection, data cleansing or cleaning is performed. There may be data that is either noisy, erroneous or missing values. Data cleaning uses different methods to eliminate this bad data from the dataset. After cleaning, data may need to be transformed as the final preparation for analytics.

In the context of smart grid, bad data detection is performed in this stage.

2.2.3 Data Representation and Modeling

Data representation and modeling are the most fundamental tasks for big data. We champion the mathematical paradigm of modeling the datasets as large random matrices, an idea first suggested in [39].

The singular value decomposition (SVD) of an arbitrary (in general complex) $p \times q$ ($p > q$) matrix $\mathbf{X}$ is given by

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^H$$

where the $p \times q$ matrix $\mathbf{U}$ has orthonormal rows, the $q \times q$ matrix $\mathbf{\Lambda}$ is diagonal with real, non-negative entries, and the $q \times q$ matrix $\mathbf{V}$ is unitary. Note that the matrices $\mathbf{X}\mathbf{X}^H = \mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^H$ and $\mathbf{X}^H\mathbf{X} = \mathbf{V}^H\mathbf{\Lambda}^2\mathbf{V}$ are Hermitian, with eigenvalues corresponding to the diagonal entries of $\mathbf{\Lambda}^2$ and $\mathbf{U}$ and $\mathbf{V}$ the corresponding matrices of eigenvectors. Consider the space-time data $I(\mathbf{x}, t)$. The SVD of such data is given by

$$I(\mathbf{x}, t) = \sum_n \lambda_n I_n(\mathbf{x}) a_n(t) \tag{2.1}$$

where $I(\mathbf{x})$ are the eigenmodes of the "spatial correlation" matrix

$$C(\mathbf{x}, \mathbf{x}') = \sum_t I(\mathbf{x}, t) I(\mathbf{x}', t)$$

and similarly $a_n(t)$ are the eigenmodes of the "temporal correlation function"

$$C(t, t') = \sum_x I(\mathbf{x}, t) I(\mathbf{x}, t')$$

We consider the case of a $p \times q$ matrix

$$\mathbf{X} = \mathbf{X}_0 + \mathbf{W}$$

where $\mathbf{X}_0$ is fixed and the entries of $\mathbf{W}$ are normally distributed with a zero mean. In general, there are correlations between entries of $\mathbf{W}$. $\mathbf{X}_0$ may be thought of as the desired or underlying "signal." For SVD to be useful, $\mathbf{X}_0$ should effectively have a low-rank structure.

$\mathbf{X}$ is a random matrix of size $p \times q$. We are interested in the large matrix limit: $p \to \infty, q \to \infty$ but the ratio $p/q \to c$.

2.2.4 Data Analysis

Our aim is to extract big data analytics. After data processing the analysis can begin. The major reason behind the need for handling big data is to be able to gain value (insight) from data analysis. Analytic techniques and methods need to be further researched to develop techniques that can process large and growing data sets. Simplification of the analysis process of big data towards an *automated* approach is a major goal behind big data.

In this stage, many different analytic methods and techniques may be performed. These methods and techniques can be broken down into three categories: statistical analysis, data mining, and machine learning. Statistical analysis creates models for predication and summarizes datasets. Data mining uses a variety of techniques (clustering, classification, etc.) to discover patterns and models present in the data. Machine learning is used to discover relationships that are present within the data.

2.2.5 Data Storage

For big data, we need an change in the architecture of systems for data storage. Data storage needs to be highly scalable and flexible enough. For storage, distributed systems like the Google File System were designed to use commodity clusters for storage. In this system, data is stored as file blocks of 64MB across the nodes of the cluster. Two additional replicas are stored to provide redundancy. On top of GFS, MapReduce is used for processing data across the nodes. It is more efficient to push computations to where the data resides rather than the opposite. MapReduce exploits the distributed architecture of the file system by sending jobs to the nodes on the cluster where the data resides.

2.3 Intelligent Algorithms

We list some promising intelligent algorithms:

- Compressive sampling, matrix completion, low-rank models, and dimensionality reduction.
- Matrix completion and low-rank matrix recovery.
- Dimensionality reduction.
- Data processing in high dimensions.
- Graph, latent factor, tensor, and multirelational data models.
- Robustness to outliers and misses; convergence and complexity issues; performance analysis.
- Scalable, online, active, decentralized, deep learning and optimization.
- Randomized schemes for very large matrix, graph, and regression problems.
- Human-machine learning systems with limited labeled and massive unlabeled data.

2.4 Signal Processing for Smart Grid

We list some promising signal processing topics for smart grid:

- Adaptive filters and statistical signal processing for smart grid.
- Distributed methods for smart grid detection, estimation, forecasting.
- Sensor fusion, data analytics, data mining, and machine learning for smart grid.
- Demand response, load management and pricing.
- Forecasting models and methods for renewable generation and for loads.
- Impacts of large-scale renewable energy integration.
- Plug in hybrid electric vehicle (PHEV) charging infrastructure and scheduling algorithms, V2G algorithms.
- Cyber-physical systems models for smart grid.
- Signal processing for smart appliances, smart meters, and sensors.

2.5 Monitoring and Optimization for Power Grids

In this chapter, by "big data" we mean smarter, more insightful data analysis. But big data is really much more than that. Companies that learn to take advantage of big data

will use *real-time* information from sensors, radio frequency identification and other identifying devices to understand their business environments:

- They pay attention to data flows as opposed to stocks.
- They rely on data scientists and product and process developers rather than data analysts.
- They are moving analytics away from the information technology (IT) function and into core business, operational, and production functions.

Recently there has been increasing interest in studying large dimensional data sets that arise in finance, wireless communications, genetics, and other fields. Patterns in these data can often be summarized by the sample covariance matrix, as done in multivariate regression and dimension reduction via factor analysis. We run the risk of being buried in the deep mathematics of summarizing big data using large random matrices (see also Section 3.1). We justify our approach by the argument that, in the infancy of big data and smart grid, the interaction between two emerging fields may be *unified* by large random matrices. This basic methodology lies at the heart of this book.

The original motivation for random matrix theory (RMT) comes from mathematical physics, where large random matrices serve as a *finite-dimensional approximation of infinite-dimensional* operators. Its importance for statistics comes from the fact that RMT may be used to correct traditional tests or estimators, which fail in the "large n, large p" setting. For example, our departure point for statistics analysis usually starts with the sample covariance matrix $\frac{1}{n}\mathbf{X}\mathbf{X}^H$. Here $\mathbf{X}$ is a complex $p \times n$ random matrix, and p and n go to infinity simultaneously, i.e., $p \to \infty, n \to \infty$ but their ratio is concentrated around c, $p/n \to c \in (0, \infty)$.

Let us first assume that the entries of $\mathbf{X}$ are i.i.d. with variance 1. Results on the global behavior of the eigenvalues of $\frac{1}{n}\mathbf{X}\mathbf{X}^H$ mostly concern the spectral distribution, $\frac{1}{p}\sum_{i=1}^{p}\delta_{\lambda_i}$, where δ denotes the Dirac measure. The spectral distribution converges, $n \to \infty, p \to \infty$ with $p/n \to c \in (0, 1]$, to a *deterministic* measure with density function

$$\frac{1}{2\pi c}\sqrt{(a-x)(b-x)}\mathbb{I}_{(a,b)}(x), \quad a = \left(1+\sqrt{c}\right)^2, b = \left(1-\sqrt{c}\right)^2$$

where $\mathbb{I}(x)$ is the indicator function. This is the so called Marchenko–Pastur law.

The remarkable observation is that when the size of random matrices are *sufficiently* large we are able to exploit a unique phenomenon: a deterministic spectral distribution is reached. The statistical properties of the entries of the large random matrix are general and flexible.

In Chapter 6, we study the large non-Hermitian random matrices in the context of free probability theory. This new framework may be useful in the context of smart grid.

2.6 Distributed Sensing and Measurement for Power Grids

A cornerstone of the smart grid is the advanced monitorability of its assets and operations. Increasingly pervasive installation of phasor measurement units (PMUs) allows so-called synchrophasor measurements to be taken roughly 100 times faster than the legacy supervisory control and data acquisition (SCADA) measurements, time stamped using the global positioning system (GPS) signals to capture the grid dynamics.

In addition, the availability of low-latency two-way communication networks will pave the way to high-precision real-time grid state estimation and detection, remedial actions to address network instability, and accurate risk analysis and post-event assessment for failure prevention. See Chapter 15 for communication and control.

Enhanced monitoring and communication capabilities lay the foundations for various grid-control and optimization components. On the distribution and consumers side, on the other hand, the demand response aims to adapt the end-user power usage in response to energy pricing, via smart metering.

For distributed generation, renewable sources such as solar, wind, and tidal, and electric vehicles are important. Based on distributed energy sources, microgrids include distributed generation and storage systems. Bidirectional power flow to/from the grid are enabled by such distributed sources. Open-grid architectures and markets are the trends.

2.7 Real-time Analysis of Streaming Data

For the smart grid, it is necessary to have appropriate methods for detailed modeling and simulation on a large scale based on the analysis of real measurements. Available consumption data are not sufficient: the coarse time-scale of measurements by so called smart meters—providing accumulated power value time series with typical frequencies of only one sample per 1 to 15 minutes—may be used for short-term local load forecasts on a statistical level but are not sufficient for global fine-grain analysis or for use in physical simulations that could increase the knowledge of dynamics and dependencies in the grid [84].

Phasor measurement units (PMUs) are high-speed sensors with the option for synchronous acquisition to enable monitoring of the power grid quality. However, PMUs are rarely used and data from the real network is expensive [85].

Electrical data recorder (EDR) measurements produce massive amounts of data every day and additional data will arise from power-grid simulations. When measuring at the high rate of 25 kHz, each EDR produces 16 GiB per day. This adds up to a total of 5.7 TiB per year per device, exceeding typical hard-disk storage sizes. As soon as we add many devices or run simulations with virtual EDRs, storing the data on disk drives connected to a single PC is obviously no longer possible and processing is not efficient. This is similar to the situation when the radio waveform data is stored—using cognitive radios as sensors [40].

Technical requirements are as follows. First, we need data storage that will not run out of space like traditional storage on hard disk. Second, we need ways to access and analyze the data efficiently. See [84] for such a system.

The model-driven approach is to (conceptually) collect or process queries on all the data sensed by the wireless sensor network (WSN). Wireless sensor networks include novel devices (e.g., smartphones), cognitive radios as sensors [39]. Sensor readings have such correlations. The model-driven approach can lead to significant energy savings for the data-acquisition task. However, due to the nature of their techniques, they can only provide probabilistic guarantees on the accuracy of the data that the sink collects, and hence no absolute bound on the error. In some scientific applications it may also be the case that the domain experts do not already have a model of the data distribution they are

sampling using the WSN but, rather, are interested in collecting accurate measurements in order to build such a model.

Consider an example of real-time content streaming. The goal is to predict the quality of service. In [86] the authors presented a new stochastic service model with capacity sharing and interruptions, appropriate for the evaluation of the quality of real-time streaming, like, for example, mobile TV, or in wireless cellular networks. The general model takes into account the multiclass Markovian process of call arrivals.

2.8 Salient Features of Big Data

Scientific advances are becoming increasingly data driven and researchers will increasingly think of themselves as consumers of data. The massive amounts of high dimensional data bring both opportunities and new challenges to data analysis. Valid statistical analysis for big data is becoming increasingly important.

In terms of computational efficiency, big data motivates the development of new computational infrastructure and data storage methods. Optimization is often a tool, not a goal, in big data analysis. Such a paradigm change has led to significant progress in the development of fast algorithms that are scalable for massive data with high dimensionality. This forges cross-fertilization among different fields including statistics, optimization, and applied mathematics.

When the data are aggregated from multiple sources, the best normalization practice remains an open problem.

Big data are characterized by massive sample size and high dimensionality. First, massive sample size allows us to unveil hidden patterns associated with small subpopulations and weak commonality across the whole population. Secondly, we discuss several unique phenomena associated with high dimensionality, including noise accumulation, spurious correlation, and incidental endogeneity. These unique features make traditional statistical procedures inappropriate.

2.8.1 Singular Value Decomposition and Random Matrix Theory

In analyzing large amounts of multivariate data, certain quantities naturally arise that are in some sense "self-averaging." Namely, in the large size limit, a single dataset can comprise a statistical ensemble for the quantity in question. One such quantity, the singular value distribution of a data matrix, is the subject of [87]. Singular value decomposition (SVD) is a representation of a general matrix of fundamental importance in linear algebra that is widely used to generate canonical representations of multivariate data. It is equivalent to principal component analysis in multivariate statistics but, in addition, is used to generate low-dimensional representations for complex multidimensional time series. One example is to generate effective low dimensional representations of high dimensional dynamical systems—called dimensionality reduction. Another example of current interest is to denoise and compress dynamic imaging data, in particular in the case of direct or indirect images of neuronal activity. Our interest in this book relates to the big data in power grids and large communications networks.

A data set with n measurements on p variables can be represented by an $n \times p$ matrix $\mathbf{X}$. In high-dimensional settings, where p is large, we often desire to reduce the

dimensionality by working with a low-rank approximation of the data matrix. The most prevalent low-rank approximation is the singular value decomposition (SVD), which is relevant to the principle component analysis (PCA). Every two or three decades, someone will claim that he invents the PCA.

Given $\mathbf{X}$, an $n \times p$ matrix, the SVD factorizes $\mathbf{X}$ as $\mathbf{X} = \mathbf{UDV}^T$, where $\mathbf{U} \in \mathbb{R}^{n \times n}$ and $\mathbf{V} \in \mathbb{R}^{p \times p}$ are orthogonal matrices and $\mathbf{D} \in \mathbb{R}^{n \times p}$ is zero except on its diagonal with diagonal entries in decreasing order. The best rank K approximation to $\mathbf{X}$, $\hat{\mathbf{X}}_K$, in both the Frobenius and operator norms, is given by the first K right singular vectors and singular values of the SVD:

$$\hat{\mathbf{X}}_K = \sum_{k=1}^{K} d_k \mathbf{u}_k \mathbf{v}_k^T.$$

In MATLAB, we have the build function *svd*. The SVD of $\mathbf{X}$ is also closely related to the eigendecomposition of $\mathbf{XX}^T$. To understand fully the implications of using the SVD in data-processing applications and classical multivariate analysis techniques such as principal components analysis (PCA), one must consider the behavior of the SVD when the elements of $\mathbf{X}$ are random.

There are two regimes of interest for random data matrices. In the first regime, the number of samples, n, is large relative to the number of variables, p, and in the second regime the two numbers are comparable. The first regime is called the "classical regime" and the second regime in the "modern" regime. The classical regime is characterized by $n \to \infty$ and p fixed; the modern regime is characterized by $n \to \infty, p \to \infty$, and $n/p \to \gamma$, where γ is a fixed scalar in $(0, \infty)$.

One can study SVD by analyzing the eigendecomposition of $\mathbf{XX}^T$. Results are stated for Gaussian random variables but many of the results hold for arbitrary distributions with finite fourth moments.

Given that the principal components direction vectors are inconsistent in high-dimensional settings, many have proposed finding principal components directions using only a subset of the variables, a method termed *sparse PCA*. This method seeks linear projections that maximize the sample variance such that these projection vectors have a limited number of nonzero elements. In other words, one seeks a direction vector $\mathbf{v}$ that maximizes var $(\mathbf{Xv})$ $/\mathbf{vv}^T$ subject to $\|\mathbf{v}\|_0 \leqslant s$, where $\|\cdot\|_0$ is the ℓ_0-norm, summing the number of nonzero elements. Jolliffe, Trendafilov and Uddin (2003) [88] first proposed estimating sparse principal components' directions by relaxing the ℓ_0-norm to an ℓ_1-norm, placing this penalty on the principal components' directions to encourage sparsity. The ℓ_1-norm is convex.

Several sparse PCA methods have been shown to be consistent in high-dimensional settings where classical PCA is inconsistent. Amini and Wainwright [89] consider a spiked covariance model.

More recently, several have proposed encouraging sparsity in both the PC directions as well as the sample principal components, forming a penalized SVD or sparse matrix factorization [90] of the following form: $\hat{\mathbf{X}}_K = \sum_{k=1}^{K} d_k \mathbf{u}_k \mathbf{v}_k^T$, where $\|\mathbf{u}_k\|_0 \leqslant t_k$, $\|\mathbf{v}_k\|_0 \leqslant s_k$.

2.8.2 Heterogeneity

Big data are often created by aggregating many data sources corresponding to different subpopulations. Each subpopulation might exhibit some unique features not shared by others.

Finite mixture models provide a flexible tool for modeling data that arise from a heterogeneous population. See [91]. Let Y be a response variable of interest and let $\mathbf{x} = \left(x_1, x_2, \ldots, x_p\right)^T$ be the vector of covariates[1] believed to have an effect on Y. We consider the mixture model for the population

$$\alpha_1 p_1\left(y; \boldsymbol{\theta}_1(\mathbf{x})\right) + \cdots + \alpha_m p_m\left(y; \boldsymbol{\theta}_m(\mathbf{x})\right) \tag{2.2}$$

where $\alpha_i \geq 0$ represents the proportion of the i-th subpopulation, $p_i\left(y; \boldsymbol{\theta}_i(\mathbf{x})\right)$ is the probability distribution of the response of the i-th subpopulation given the covariates $\mathbf{x}$, $\boldsymbol{\theta}_i(\mathbf{x})$ as the parameter vector. In practice, many subpopulations are rarely observed, so α_i is very small. Because big data are characterized by large sample size n, the sample size $n\alpha_i$ for the i-th subpopulation can be moderately large even if α_i is very small.

Inferring the mixture model in (2.2) for large datasets requires sophisticated statistical and computational methods. In high dimensions, however, we need to regularize the estimating procedure carefully to avoid overfitting or noise accumulation [92, 93]. The authors in [94] proposed an ℓ_1-regularized likelihood method for estimating the inverse covariance matrix in the high-dimensional multivariate normal model in presence of *missing* data. Their method is based on the assumption that the data are missing at random.

2.8.3 Noise Accumulation

Analyzing big data requires us simultaneously to estimate or test many parameters. These estimate errors accumulate when a decision or prediction rule depends on a large number of such parameters. Such a noise-accumulation effect is especially severe in high dimensions and may even dominate the true signals. It is usually handled by the sparsity assumption [95–97].

Consider a classification problem [98] where the data come from two classes

$$\mathbf{X}_1, \ldots, \mathbf{X}_n \sim \mathcal{N}\left(\boldsymbol{\mu}_1, \mathbf{I}_d\right) \quad \text{and} \quad \mathbf{Y}_1, \ldots, \mathbf{Y}_n \sim \mathcal{N}\left(\boldsymbol{\mu}_1, \mathbf{I}_d\right) \tag{2.3}$$

We want to construct a classification rule that classifies a new observation $\mathbf{Z} \in \mathbb{R}^d$ into either the first or the second class. For example, for $n = 100$ and $d = 1000$, we set $\boldsymbol{\mu}_1 = \mathbf{0}$ and $\boldsymbol{\mu}_2$ to be sparse: only the first ten entries of $\boldsymbol{\mu}_2$ is nonzero with value 3, all the other entries are zero. When $m = 2$, we can get high discriminative power. However, the discriminative power becomes very low when m is too large due to noise accumulation. The first ten features contribute to classifications and the remaining features do not. Therefore, when $m > 10$, procedures do not get any additional signals but accumulate noise: the larger m, the more noise accumulation, which makes the classification procedure deteriorate with dimensionality.

2.8.4 Spurious Correlation

High dimensionality also brings spurious correlation, which means that many uncorrelated random variables may have high sample correlations in high dimensions. Spurious correlation may lead to wrong statistical inferences.

Consider the problem of estimating the coefficient vector of a linear model

$$\mathbf{y} = \mathbf{X}\mathbf{z} + \mathbf{w}, \quad \text{Var}(\mathbf{w}) = \sigma^2 \mathbf{I}_d \tag{2.4}$$

1 In statistics, a covariate is a variable that is possibly predictive of the outcome under study.

where $\mathbf{y} \in \mathbb{R}^n$ represents the response vector, $\mathbf{X} = \left[\mathbf{x}_1, \ldots \mathbf{x}_n\right]^T \in \mathbb{R}^{n \times d}$ represents the design matrix, $\mathbf{w} \in \mathbb{R}^n$ represents an independent random noise vector, and $\mathbf{I}_d$ is the $d \times d$ identity matrix.

In high dimensions, even for a model as simple as (2.4), variable selection is challenging due to the presence of spurious correlation. When the dimensionality is high, the important variables can be highly correlated with several spurious variables that are scientifically unrelated [99].

2.8.5 Incidental Endogeneity

Incidental endogeneity is another subtle issue raised by high dimensionality. In a regression setting $Y = \sum_{i=1}^{d} \beta_i X_i + W$, the term "endogeneity" means that some predictors X_i correlate with the residual noise W. The conventional sparse model assumes

$$Y = \sum_{i=1}^{d} \beta_i X_i + W, \quad \text{and} \quad \mathbb{E}\left(WX_i\right) = 0 \quad \text{for } i = 1, \ldots, d \tag{2.5}$$

with a small set $S = \left\{i : \beta_i \neq 0\right\}$. The exogenous assumption in (2.5) that the residual noise W is uncorrelated with all the predictors is crucial for the validity of most existing statistical procedures, including variable selection consistency. Though this assumption looks natural, it is easy to be violated in high dimensions as some of variables X_i are incidentally correlated with W, making most high-dimensional procedures statistically invalid.

2.8.6 Impact on Computational Methods

Big data are massive and very high dimensional, which poses significant challenges on computing and paradigm shifts on large-scale optimization [100]. Direct application of penalized quasilikelihood estimators on high-dimensional data requires us to solve very large-scale optimization problems. Parallel computing, randomized algorithms, approximate algorithms, and simplified implementations are promising. See [40] for randomized algorithms.

The volumes of modern datasets are exploding and it is often computationally infeasible to make inferences directly based on the raw data. As a result, to handle big data from both a statistical perspective and a computational perspective, dimension reduction as a data preprocessing step is exploited [101].

2.9 Big Data for Quantum Systems

In this book, the first big data system is traced back to large quantum systems in the 1950s.

For a large class of quantum systems, the statistical properties of their spectrum show remarkable agreement with random matrix predictions. Recent advances show that the scope of random matrix theory is much wider.

The study of random matrix ensembles has provided deep insight in several fields of physics including nuclear, atomic and molecular physics, quantum chaos, and mesoscopic systems. The interest in random matrices arose from the need to understand the

spectral properties of the many-body quantum systems with complex interactions—the challenge of the data deluge for the first time in 1950s. With general assumptions about the symmetry properties of the system dictated by quantum physics, random matrix theory (RMT) provides remarkably successful *predictions* for the statistical properties of the spectrum.

The fluctuation properties of low-dimensional systems, such as chaotic quantum systems, are universal and can be modeled by an appropriate ensemble of random matrices. Random matrix techniques have potential applications and utility in disciplines far outside of quantum physics.

2.10 Big Data for Financial Systems

For big data, data capture, data storage and data representation are fundamental. Big data analytics can be extracted from big data. The aim of this section is to introduce how financial data is represented—data modeling. Once the data is appropriately represented, the rest of problems relate to how to analyze the data to extract useful information (or knowledge). Financial data are very common so we use the financial system as the prototype for other datasets. Another reason is that a lot of research has been done in the financial literature so we can apply, by the method of analogy, the massive results available there to our problems at hand, such as power grids, sensing networks, and communication networks.

2.10.1 Methodology

It was the economy which followed physics, and not vice versa. The father of classical economics, Adam Smith, exemplifies the methodology of science by stressing the role of observing the regularities and then constructing theories (which Smith called "imaginary machines") reproducing the observations. Using astronomy as a reference point was not accidental—it was the celestial mechanics, and the *impressive amount of astronomical data*, which dominated science in several cultures.

One of the benefits of computers was that economic systems started to *save* more and more data. Today markets collect incredible amounts of data (they remember practically every transaction). This triggers the need for new methodologies, able to manage the data. In particular, the data started to be analyzed using methods borrowed widely from physics, where seeking regularities for unconventional correlations is mandatory. In the new science of big data, financial engineers are ahead of communications engineers because their data are more accessible.

Since the mid-1990s there has been a trend—physicists started to study the economy scientifically. These studies were devoted mostly to quantitative finance. To a large extent, it was triggered by vast amount of data accessible in this field—big data. In this way, physics started to play the role of financial mathematics—sometimes rephrasing the mathematical constructions in the language of physics, sometimes applying methods developed solely in physics, usually at the level of various effective theories of complex systems.

The aim of macroeconomic studies is to extract important factors, understand their mutual relations and describe the development of past events. The ultimate goal is to

reach a level of understanding that would also permit the *prediction* of the system's reaction to changes in macroeconomic parameters in the future.

In some ways, the problem of interpreting the correlations between individual stock-price changes—and also the problem of data deluge in the age of big data—is reminiscent of the difficulties experienced by physicists in the 1950s, in interpreting the spectra of complex nuclei. Large amounts of spectroscopic data on the energy levels were becoming available but were too complex to be explained by model calculations because the exact nature of the interactions was unknown. Random matrix theory (RMT) was developed in this context to deal with the statistics of energy levels of complex quantum systems [102]. With the *minimal assumption* of a random Hamiltonian, given by a real symmetric matrix with independent random elements, a series of remarkable predictions was made and successfully tested on the spectra of complex nuclei [102]. Indeed, they postulated that the Hamiltonian describing a heavy nucleus could be described by a matrix **H** with independent random elements H_{ij} drawn from a probability distribution. Random matrix theory predictions represent an average over all possible interactions [102]. Deviations from the universal predictions of RMT—anomalies detection—identify *system-specific*, *nonrandom* properties of the system under consideration, providing clues about the nature of the underlying interactions [103–105]. Random matrix techniques have potential applications and utility in disciplines far outside of quantum physics.

Quantifying correlations between different stocks is a topic of interest not only for scientific reasons of understanding the economy as a complex dynamical system but also for practical reasons such as asset allocation and portfolio-risk estimation. Unlike most physical systems, where one relates correlations between subunits to basic interactions, the underlying "interactions" for the stock market problem are *not known*. Here, we analyze cross correlations between stocks by applying concepts and methods of random matrix theory, developed in the context of complex quantum systems where the precise nature of the interactions between subunits are *not known*. By analogy, we extend this theory to the general big data systems, where the precise nature of the interactions between subsystems are *not known*, such as power grids, sensor networks [40], large communication systems (massive MIMO and cognitive radio network [39]), and even atmospheric correlations [106].

Based on the reasoning above, one may trace the big data problems back to complex quantum systems in the 1950s. The unified idea of this book is to model big data as large random matrices so we can use RMT to extract big data analytics. The underlying assumption is that RMT is firmly established in physics after 60 years' research. In this sense, we place RMT at the heart of the theory for big data analytics. We emphasize the universality of RMT so we can apply RMT to a huge class of big data problems. To make the point, our principle is that we only consider those big data problems whose data can be represented by large random matrices. We follow this principle in cognitive radio network [39] and cognitive sensing [40]. It is reasonable to extend this principle to other large datasets such as transportation and manufacturing.

Below, we apply RMT methods to study the cross correlations of stock-price changes, following [107]. We consider N assets; the correlation matrix contains $N(N-1)/2$ entries, which must be determined from N time series of length T; if T is not very large compared to N, one should expect that the determination of the covariances is noisy, and therefore that the empirical correlation matrix is, to a large extent,

random—that the structure of the matrix is dominated by measurement noise. If this is the case, one should be very careful when using this correlation matrix in applications. In particular, the smallest eigenvalues of this matrix are the most sensitive to this "noise". It is thus important to devise methods that allow one to distinguish "signal" from "noise",[2]—eigenvectors and eigenvalues of the correlation matrix containing real information from those that are devoid of any useful information, and, as such, unstable in time. From this point of view, it is interesting to compare the properties of an empirical correlation matrix **C** to a "null hypothesis" purely random matrix that one could obtain from a finite time series of strictly uncorrelated assets. Deviations from the random matrix case might then suggest the presence of true information.

Recent studies applying RMT methods to analyze the properties of C show that $\approx 98\%$ of the eigenvalues of **C** agree with RMT predictions, suggesting a considerable degree of randomness in the measured cross correlations. It was also found that there are deviations from RMT predictions for $\approx 2\%$ of the largest eigenvalues. These results prompt the following questions:

- What is a possible interpretation of the deviations from RMT?
- Are the deviations from RMT stable in time?
- What can we infer about the structure of **C** from these results?
- What are the practical implications of these results?

Initially, RMT was proposed to explain energy spectra of complicated nuclei half a century ago. In its simplest form, a random matrix ensemble is an ensemble of $N \times N$ matrices **A** whose entries A_{ij} are uncorrelated i.i.d. random variables, and whose distribution is given by

$$\mathbb{P}(\mathbf{A}) \sim \exp\left(-\frac{\beta N}{2}\,\mathrm{Tr}\left(\mathbf{A}\mathbf{A}^T\right)\right) \tag{2.6}$$

where β takes specific values for different ensembles of matrices (e.g. depending on whether or not the random variables are complex or real valued). Eigenvalue spectra and correlations of eigenvalues in the limit $\mathbf{N} \to \infty$ have been worked out for symmetric $N \times N$ random matrices by Wigner [109, 110]. For real valued matrix entries, such symmetric random matrices are sometimes referred to as the Gaussian orthogonal ensemble (GOE).

The symmetry constraint has later been relaxed by Ginibre and the probability distributions of different ensembles (real, complex, quaternion)—known as Ginibre ensembles (GinOE, GinUE, GinSE)—have been derived [111] in the limit of infinite matrix size. For ensembles of random real asymmetric matrices (GinOE)—the most difficult case—progress has only slowly been made with great effort in recent decades. The eigenvalue density could finally be derived via different methods [112, 113], where—quite remarkably—the finite-size dependence of the ensemble has also been elucidated [113]. For recent progress in the field see [114].

Biely and Thurner (2008) [115], for the first time in financial applications, applied (lagged) covariance matrices stemming from finite rectangular $N \times T$ data matrices **X**, which contain data for N different assets (or instruments) at T observation points. The matrix ensemble corresponding to the $N \times N$ covariance matrix $\mathbf{C} \sim \mathbf{X}\mathbf{X}^T$ of such

2 The central task of big data analytics is to distinguish "signal" from "noise" [108].

data is known as the Wishart ensemble [116] and is a *cornerstone* of multivariate data analysis. For the case of uncorrelated Gaussian distributed data, the exact solution to the eigenvalue-spectrum of $\mathbf{X}\mathbf{X}^T$ is known as Marchenko–Pastur law (for $N \to \infty$) and has been used as a *starting point* for random matrix analysis of correlation matrices at lag zero. Besides, a quite general methodology of extracting meaningful correlations between variables has been discussed based on a generalization of the Marcenko–Pastur distribution [82]. The underlying method was the powerful tool of singular-value decomposition and RMT was used to predict singular-value spectra of Gaussian randomness.

The time-lagged analogon to the covariance matrix is defined as

$$C_\tau^{ij} \sim \sum_{t=1}^{T} X_t^i X_{t-\tau}^j$$

where one time series is shifted by τ timesteps with respect to the other. In contrast to (real-valued) equal-time correlation matrices of the Wishart ensemble, which have a real eigenvalue spectrum, the spectrum of $\mathbf{C}_\tau$ is defined in the complex plane, as matrices of this type are, in general, asymmetric. It is the analysis of the asymmetric time-lagged correlations that forms a fundamental part of finance and econometrics.

2.10.2 Marchenko–Pastur Law for Equal Time Correlations

From the point of view of noncommutative probability and central limit theorems, the result of this subsection is natural and fundamental. From this point of view, it is also puzzling how late the random matrices (in our language matrix probabilities) were used for the analysis of financial data. The breakthrough came in 1999 [107, 117].

The empirical correlation matrix $\mathbf{C}$ is constructed from the time series of price changes $X_i(t)$ (where $i = 1, \ldots, N$ labels the asset and $t = 1, \ldots, T$ the time) through the equation:

$$C_{ij} = \frac{1}{T} \sum_{t=1}^{T} X_i(t) X_j(t) \tag{2.7}$$

We can symbolically write (2.7) as

$$\mathbf{C} = \frac{1}{T} \mathbf{X}\mathbf{X}^T \tag{2.8}$$

where $\mathbf{X}$ is an $N \times T$ rectangular matrix, and T denotes matrix transposition. The null hypothesis of uncorrelated assets, which we consider now, translates itself in the assumption that the coefficients $(\mathbf{X})_{it} = X_i(t)$ are independent, identically distributed, random variables. We denote $\rho_{\mathbf{C}}(\lambda)$ the density of eigenvalues of $\mathbf{C}$, defined as:

$$\rho_{\mathbf{C}}(\lambda) = \frac{1}{N} \frac{dn(\lambda)}{d\lambda} \tag{2.9}$$

where $n(\lambda)$ is the number of eigenvalues of $\mathbf{C}$ less than λ. Interestingly, if $\mathbf{X}$ is a $T \times N$ random matrix, $\rho_{\mathbf{C}}(\lambda)$ is exactly known in the limit $N \to \infty$, $T \to \infty$ and $c = T/N \geq 1$ fixed, and follows the so-called Marchenko–Pastur law:

$$\rho_{\mathbf{C}}(x) = \frac{c}{2\pi\sigma^2} \frac{\sqrt{(b-x)(x-a)}}{x} \tag{2.10}$$

where

$$a = \sigma^2 \left(1 + 1/c - 2\sqrt{1/c}\right), \quad b = \sigma^2 \left(1 + 1/c + 2\sqrt{1/c}\right),$$

with $x \in [a, b]$ where σ^2 is equal to the variance of the elements of $\mathbf{X}$, equal to 1 with our normalization. In the limit $c = 1$ the normalized eigenvalue density of the matrix $\mathbf{X}$ is the well known Wigner semicircle law, and the corresponding distribution of the *squares* of these eigenvalues. The most important features predicted by (2.10) are:

- the fact that the lower "edge" of the spectrum is strictly positive (except for $c = 1$); there are therefore no eigenvalues between 0 and a. Near this edge, the density of eigenvalues exhibits a sharp maximum, except in the limit $c = 1$ ($a = 0$) where it diverges as $\sim 1/\sqrt{x}$;
- the density of eigenvalues also vanishes above a certain upper edge b.

Note that the above results are only valid in the limit $N \to \infty$. For finite N, the singularities present at both edges are smoothed: the edges become somewhat blurred, with a small probability of finding eigenvalues above b and below a, which goes to zero when N becomes large.

Now, we want to compare the empirical distribution of the eigenvalues of the correlation matrix of stocks corresponding to different markets with the theoretical prediction given by (2.10), based on the assumption that the correlation matrix is random. We have studied numerically the density of eigenvalues of the correlation matrix of $N = 406$ assets of the S&P 500, based on daily variations during the years 1991–96, for a total of $T = 1309$ days (the corresponding value of $c = T/N$ is 3.22).

The unexpected results showed that the majority of the spectrum of empirical covariance matrices is populated by noise! Only a few of the largest eigenvalues did not match the pattern. An immediate observation is that the highest eigenvalue λ_1 is 25 times larger than the predicted b—see Figure 2.1, inset. The simplest "pure noise" hypothesis is therefore inconsistent with the value of λ_1. A more reasonable idea is that the components of the correlation matrix which are orthogonal to the "market" is pure noise. One can treat σ^2 as an adjustable parameter. The best fit is obtained, for example, using the least-square method, for $\sigma^2 = 0.74$, and corresponds to the dark line in Figure 2.1, which accounts quite satisfactorily for 94% of the spectrum, while the 6% highest eigenvalues still exceed the theoretical upper edge by a substantial amount. Note that a still better fit could be obtained by allowing for a slightly smaller effective value of c, which could account for the existence of volatility correlations.

2.10.3 Symmetrized Time-Lagged Correlation Matrices

The previous methods in Section 2.10.2 involved equal time correlations. The construction of delay correlation matrix involves calculating correlations between different entities with a time delay. Consider the multivariable time series at hand represented as a matrix $\mathbf{X}$ of order $N \times T$. Here N is the number of time series of length T each. Suppose that i and j are two time series among the given multivariable time series $\mathbf{X}$. The correlation between i at say $t = 0$ and j at time lag $t = \tau$ is given by

$$C_{ij} = \frac{1}{T} \sum_{t=1}^{T} X_{it} X_{j(t+\tau)} \tag{2.11}$$

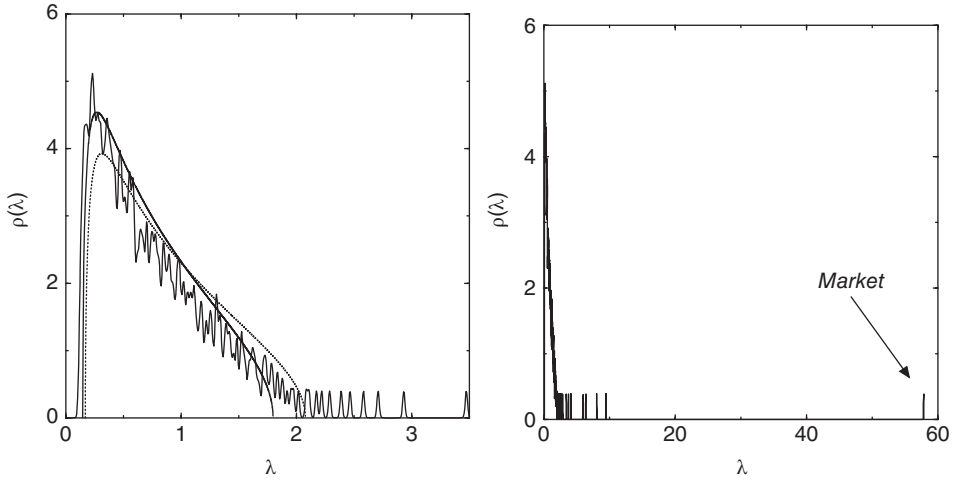


Figure 2.1 Smoothed density of the eigenvalues of $\mathbf{C}$, where the correlation matrix $\mathbf{C}$ is extracted from $N = 406$ assets of the S&P500 during the years 1991–1996. For comparison we have plotted the density (2.10) for $c = 3.22$ and $\sigma^2 = 0.85$: this is the theoretical value obtained assuming that the matrix is purely random except for its highest eigenvalue (dotted line). A better fit can be obtained with a smaller value of $\sigma^2 = 0.74$ (solid line), corresponding to 74% of the total variance. Inset: same plot, but including the highest eigenvalue corresponding to the "market," which is found to be 30 times greater than b. Source: Reproduced with permission from [107].

The matrix $\mathbf{C}$ thus constructed is asymmetric. The eigenvalues of such a matrix will be complex. To have real eigenvalues we suitably symmetrize the matrix $\mathbf{C}$. The symmetrized matrix $\mathbf{C}^S$ is constructed according to the expression

$$C_{ij}^S = C_{ji}^S = \frac{C_{ij} + C_{ji}}{2}$$

The matrix element X_{it} corresponds to the t-th element of the time series i. The symmetrized delay correlation matrix $\mathbf{C}^S$ may be thus represented in terms of the matrix $\mathbf{X}$ by the expression

$$\mathbf{C}^S = \frac{\mathbf{X}^H(0)\,\mathbf{X}(\tau) + \mathbf{X}(\tau)\,\mathbf{X}(0)}{2T}$$

We see [118] for empirical data sets of atmospheric data and stock market data. They construct such matrices for varying delay values.

For the case of the independent, identically distributed data matrix $\mathbf{X}$, the resolvent is defined as

$$G(z) = \mathrm{Tr}\left(\left(z - \mathbf{X}^H\mathbf{X}\right)^{-1}\right)$$

where $G(z)$ is a complex function. The density of eigenvalues [87] is given by

$$\rho(\lambda) = \sum_i \delta\left(\lambda - \lambda_i\right) = \frac{1}{\pi}\lim_{\varepsilon\to\infty}\mathrm{Im}\left[G(\lambda - i\varepsilon)\right] \tag{2.12}$$

The above expression is valid for a simple correlation matrix for independent, identically distributed random variables. For the symmetrized delay correlation matrix $\mathbf{C}^S$, the derivation is done by using the resolvent in

$$G_\tau(z) = \mathrm{Tr}\left(\left(z - \mathbf{C}^S(\tau)\right)^{-1}\right) \tag{2.13}$$

The expression for $G_\tau(z)$ is obtained by expanding the resolvent in powers of $1/z$ and using a diagrammatic technique to represent various terms in the expansion. In the limit $N, T \to \infty$, while maintaining the ratio $Q = (T - \tau)/N$ to be a constant, only the planar diagrams contribute [87]. We can sum the diagrams to infinite order and, for $\tau << N$, we obtain the following fourth-order equation for $G_\tau(z)$

$$G^4 + 2\kappa G^3 + \left(\kappa^2 - \frac{Q^2}{\sigma^4}\right) G^2 - 2\kappa \frac{Q^2}{\sigma^4} G + \frac{Q^2}{\sigma^4 \lambda^2}(2Q - 1) = 0 \tag{2.14}$$

where $G_\tau(z)$ is represented by G for convenience and $\kappa = \frac{Q-1}{\lambda}$

Equation (2.14) is solved numerically to obtain the required solution for G. The imaginary part of the solution for G is substituted in (2.12) to obtain the eigenvalue distribution for the delay correlation matrix. See Figure 2.2 for illustration. The analytical model agrees with the numerical simulations, as shown in Figure 2.3.

2.10.4 Asymmetric Time-Lagged Correlation Matrices

The entries in the $N \times T$ data matrices $\mathbf{X}$ for N assets and T observation times, are the log-return time-series of asset i at observation times t,

$$X_t^i = \ln S_t^i - \ln S_{t-1}^i \tag{2.15}$$

after subtraction of the mean and normalization to unit variance, i.e. division by $\sigma_i = \sqrt{\left\langle \left(X_t^i\right)^2 \right\rangle - \left(X_t^i\right)^2}$. Here, S_t^i is the price of asset i at time t. One time unit is the time difference between observations at $t + 1$ and t, for example a day, 5 minutes; for tic

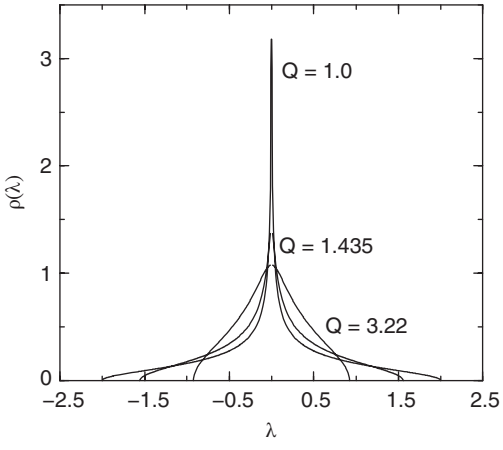


Figure 2.2 Plots of $\rho(\lambda)$ versus for different values of $Q = 1$, $Q = 1.435$ and $Q = 3.22$.

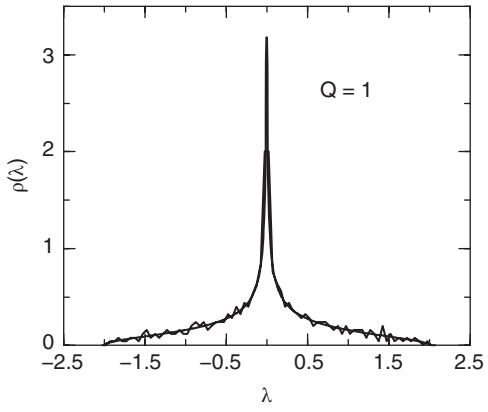


Figure 2.3 Combined plot of $\rho(\lambda)$ versus λ for $Q = 1$ obtained analytically as well as numerically for independent, identically distributed random data sets.

data it can also be of variable size. Time-lagged correlation functions of unit-variance log-return series among stocks are defined as

$$C_{\tau}^{ij}(T) \equiv \left\langle \left(X_t^i - \left\langle X_t^i \right\rangle\right)\left(X_{t-\tau}^i - \left\langle X_{t-\tau}^i \right\rangle\right)\right\rangle_T \tag{2.16}$$

where the time-lag τ is measured in time units and $< \cdots >_T$ stands for a time-average over the period T. We drop T below. Equal-time correlations are obviously obtained for $\mathbf{T} = 0$. For $\tau \neq 0$, the lagged correlation matrix $\mathbf{C}_\tau$ is generally not symmetric and contains the lagged auto-correlations in the diagonal. It can be written as

$$\mathbf{C}_\tau = \frac{1}{T}\mathbf{X}\mathbf{D}_\tau\mathbf{X}^T \tag{2.17}$$

where $\mathbf{D}_\tau \equiv \delta_{t,t+\tau}$ and where $\mathbf{X}$ is the $N \times T$ normalized time-series data. Denoting the eigenvalues of (C_τ^{ij}) by λ_i and their associated eigenvectors by $\mathbf{u}_i$ (or u_{ik}), where $i, k = 1, \ldots, N$, we may write the eigenvalue problem as

$$\sum_j C_\tau^{ij}\mathbf{u}_i = \lambda_j\mathbf{u}_j \tag{2.18}$$

We immediately recognize that eigenvalues λ_i are either real or complex conjugates, because the matrix elements of C_τ^{ij} are real and thus the conjugate eigenvalue λ_i^* also solves (2.18). Regarding the elements of C_τ^{ij} as random variables with a certain distribution, we should keep in mind that their specific construction, (2.17), results in a departure from a "purely" random real asymmetric $N \times N$ matrix where the entries are i.i.d. Gaussian distributed.

2.10.5 Noise Reduction

Consider the empirical covariance matrix defined in (2.8), repeated here for convenience:

$$\mathbf{S} = \frac{1}{T}\mathbf{X}\mathbf{X}^T \tag{2.19}$$

which is an $N \times N$ matrix.

The comparison with random matrix theory proved not only helpful to identify the noise in correlation matrices but also showed a way to reduce this noise. The RMT

filtering method was developed in [119]. After diagonalization of the correlation matrix,

$$\mathbf{\Lambda} = \mathbf{USU}^{-1}$$

only $N - s$ highest eigenvalues are preserved, while the bulk of the eigenvalues are set to zero. This filtered spectrum,

$$\mathbf{\Lambda}^{\text{filtered}} = \text{diag}\left(0, \ldots, 0, \lambda_{s+1}, \ldots, \lambda_N\right)$$

is then transformed back into the original basis:

$$\mathbf{S}^{\text{filtered}} = \mathbf{U}^{-1}\mathbf{\Lambda}^{\text{filtered}}\mathbf{U}$$

Finally, the normalization of the diagonal to 1 has to be restored: $S_{ii}^{\text{filtered}} = 1$ for all i.

This method is capable of removing the noise for uncorrelated assets completely. While keeping only the significant eigenvalues, the information about the full correlation structure and also the noise are still present in the eigenvectors. They are included in the unitary matrix $\mathbf{U}$ which is used to transform back to the original basis.

A weakness of the RMT filtering method is that it discards information buried in the bulk of the spectrum. This can become relevant for correlation structures with many small and weakly correlated branches. A method that avoids such a cutoff has been introduced by Guhr and Kalber (2003) [120]—the so-called power mapping. It takes each element of the correlation matrix and raises its absolute value to some power q, while preserving the sign,

$$C_{ij}^{(q)} = \text{sign}\left(C_{ij}\right)\left|C_{ij}\right|^q \tag{2.20}$$

It is worth pointing out that $C^{(q)}$ is not the same as C^q. When q is larger than 1, the entries in the matrix will be suppressed, because, due to normalization, their absolute values are smaller than or equal to 1. The idea behind power mapping is that the noise will be suppressed more strongly than the actual correlations. This can be seen, for example, in the spectral density. The power mapping has a similar effect on the spectral density as a prolongation of the time series. However, if q becomes too large, the actual correlations will be suppressed more and more.

The power-mapping method is reminiscent of the power of the non-Hermitian random matrix $\mathbf{X}$ in Section 6.6, $\mathbf{X}^\alpha$, for an arbitrary real number α.

2.10.6 Power-Law Tails

To what extent the "historical" determination of covariance estimators (i.e. based on past time series over a finite temporal window T) can be trusted when forecasting the financial risk of a certain portfolio; put differently, how reliable is the past in shaping the future? In a pioneering paper, Laloux *et al.* [107] used a comparison with RMT to cast serious doubts on the usefulness of historical covariance spectra in estimating the variance of a given portfolio, questioning the widely applied procedure of Markowitz's theory based on Gaussian mean-field approximations. The "measurement noise" due to the *finiteness* of the historical time series T was claimed in [107] to bury most of the relevant information encoded in the historical covariance matrices, thus impairing, from the beginning, much of the consequent predictions. Clever methods were devised

to detect meaningful correlations buried under the "noise-dressed" regions of the spectra [120, 121] thus trying to mitigate the pessimistic forecast of [107].

Consider a statistical system with N correlated random variables. Imagine that we do not know *a priori* correlations between the variables and that we try to learn about them by sampling the system T times. Results of the sampling can be stored in a rectangular matrix $\mathbf{X}$ containing empirical data X_{it}, where the indices $i = 1, \ldots, N$ and $t = 1, \ldots T$ run over the set of random variables and measurements, respectively. If the measurements are uncorrelated in time the two-point correlation function reads

$$\left\langle X_{i_1 t_1} X_{i_2 t_2} \right\rangle = C_{i_1 i_2} \delta_{t_1 t_2} \tag{2.21}$$

where $\mathbf{C}$ is called correlation matrix or covariance matrix. For simplicity assume that $\langle X_{it} \rangle = 0$. If one does not know $\mathbf{C}$, one can try to reconstruct it from the data $\mathbf{X}$ using the empirical covariance matrix

$$C_{ij} = \frac{1}{T} \sum_{t=1}^{T} X_{it} X_{jt} \tag{2.22}$$

which is a standard estimator of the correlation matrix. One can think of $\mathbf{X}$ as of an $N \times T$ random matrix chosen from the matrix ensemble with some prescribed probability measure $\mathbb{P}(\mathbf{X}) d\mathbf{X}$. The empirical covariance matrix:

$$\mathbf{S} = \frac{1}{T} \mathbf{X} \mathbf{X}^T \tag{2.23}$$

thus depends on $\mathbf{X}$. For the given random matrix $\mathbf{X}$, the eigenvalue density of the empirical matrix $\mathbf{S}$ is

$$\rho(\mathbf{X}, \lambda) \equiv \frac{1}{N} \sum_{i=1}^{N} \delta\left(\lambda - \lambda_i(\mathbf{S})\right) \tag{2.24}$$

where $\lambda_i(\mathbf{S})$ denotes eigenvalues of $\mathbf{S}$. Averaging over all random matrices $\mathbf{X}$

$$\rho(\mathbf{X}, \lambda) \equiv \langle \rho(\mathbf{X}, \lambda) \rangle = \int \rho(\mathbf{X}, \lambda) \mathbb{P}(\mathbf{X}) D\mathbf{X} \tag{2.25}$$

we can find the eigenvalue density of $\mathbf{S}$ which is representative for the whole ensemble of $\mathbf{X}$. We are interested in how the eigenvalue spectrum of $\mathbf{S}$ is related to that of $\mathbf{C}$.

The question is how to clean the spectrum of the empirical matrix $\mathbf{S}$ from the noise optimally in order to obtain a best quality estimate of the spectrum of the underlying exact covariance matrix $\mathbf{C}$. One can consider a more general problem, where in addition to the correlations between the degrees of freedom (stocks) there are also temporal correlations between measurements [122]

$$\left\langle X_{i_1 t_1} X_{i_2 t_2} \right\rangle = C_{i_1 i_2} A_{t_1 t_2} \tag{2.26}$$

given by an autocorrelation matrix $\mathbf{A}$. If $\mathbf{X}$ is a Gaussian random matrix, or more precisely if the probability measure $\mathbb{P}(\mathbf{X}) D\mathbf{X}$ is Gaussian, then the problem is analytically solvable in the limit of large matrices [87, 121–123]. One can then derive an *exact* relation between the eigenvalue spectrum of the empirical covariance matrix $\mathbf{S}$ and the spectra of the correlation matrices $\mathbf{A}$ and $\mathbf{C}$.

There is a model that, on the one hand, keeps the structure of correlations (2.26) and, on the other hand, has power-law tails in the marginal probability distributions for individual matrix elements. More generally, we will calculate the eigenvalue density of the

empirical covariance matrix $\mathbf{S}$ (2.23) for random matrices $\mathbf{X}$ which have a probability distribution of the form

$$\mathbb{P}(\mathbf{X})\, D\mathbf{X} = \mathcal{N}^{-1} f\left(\operatorname{Tr} \mathbf{X}^T \mathbf{C}^{-1} \mathbf{X} \mathbf{A}^{-1}\right) D\mathbf{X} \tag{2.27}$$

where $D\mathbf{X} = \prod_{i,t=1}^{N,T} dX_{it}$ is a volume element. The normalization constant $\mathcal{N}$

$$\mathcal{N} = \pi^{d/2} (\det \mathbf{C})^{T/2} (\det \mathbf{A})^{N/2} \tag{2.28}$$

and the parameter $d = NT$ have been introduced for convenience. The function f is an arbitrary non-negative function such that $\mathbb{P}(\mathbf{X})$ is normalized: $\int \mathbb{P}(\mathbf{X})\, D\mathbf{X} = 1$.

In particular we will consider an ensemble of random matrices with the probability measure given by a multivariate Student distribution

$$\mathbb{P}(\mathbf{X})\, D\mathbf{X} = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\mathcal{N}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \operatorname{Tr} \mathbf{X}^T \mathbf{C}^{-1} \mathbf{X} \mathbf{A}^{-1}\right)^{-(\nu+d)/2} D\mathbf{X} \tag{2.29}$$

The two-point correlation function can be easily calculated for this measure:

$$\left\langle X_{i_1 t_1} X_{i_2 t_2} \right\rangle = \frac{\sigma^2}{\nu - 2} C_{i_1 i_2} A_{t_1 t_2}$$

We see that for $\sigma^2 = \nu - 2$ and for $\nu > 2$ the last equation takes the form (2.26).

Let us first consider the case without correlations: $\mathbf{C} = \mathbf{I}_N$ and $\mathbf{A} = \mathbf{I}_T$. The spectrum of the empirical covariance for the Gaussian ensemble is given by the Marchenko–Pastur distribution:

$$\rho_G(\lambda) = \frac{1}{2\pi c \lambda} \sqrt{(b - \lambda)(\lambda - a)}$$

where $a = \left(1 - \sqrt{c}\right)^2$, and $b = \left(1 + \sqrt{c}\right)^2$.

The corresponding spectrum for the Student ensemble is then

$$\rho_\nu(\lambda) = \frac{1}{2\pi c \Gamma\left(\frac{\nu}{2}\right)} \left(\frac{\nu}{2}\right)^{\nu/2} \lambda^{-\nu/2-1} \int_a^b \sqrt{(b - x)(x - a)} e^{-\nu x/2\lambda} x^{(\nu/2)-1} dx \tag{2.30}$$

The integral over dx can be easily computed numerically. Results of this computation for different values are shown in Figure 2.4 and Figure 2.5.

2.10.7 Free Random Variables

The law of large numbers and the central limit theorem are two cornerstones of the theory of probability. Understanding their relationships with the physical laws is central to all the science. Large random matrices can be regarded as free random variables. This approach is basic to many results published in physics and finance. See [125] for its application in financial data.

The concepts of the free random variables (FRV) calculus serve as a powerful alternative to standard random matrix theory, both for Gaussian and non-Gaussian noise. Free random variables may be thought of as an abstract noncommutative generalization of the classical (commutative) probability calculus, i.e. a mathematical framework

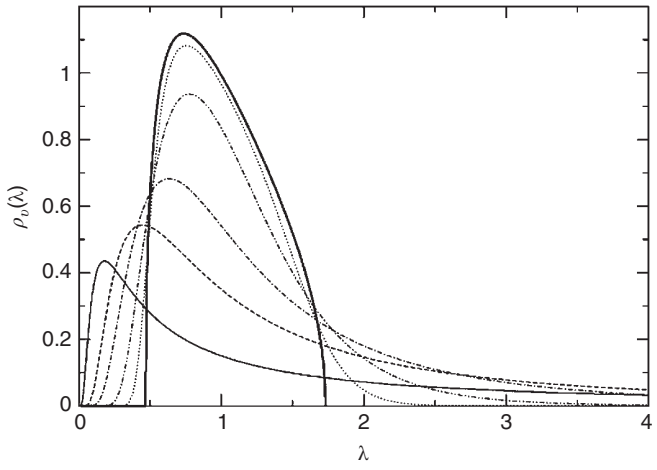


Figure 2.4 Spectra of the covariance matrix **C** for the Student distribution (2.29) with $\mathbf{C} = \mathbf{I}_N$ and $\mathbf{A} = \mathbf{I}_T$, $c = N/T = 0.1$, for $\nu = 1/2$, 2, 5, 20, and 100 (thin lines from solid to dotted), calculated using the formula (2.30) and compared to the uncorrelated Wishart (thick line). One sees that for $\nu \to \infty$ the spectra tend to the Wishart distribution. Source: Reproduced from [124] with permission.

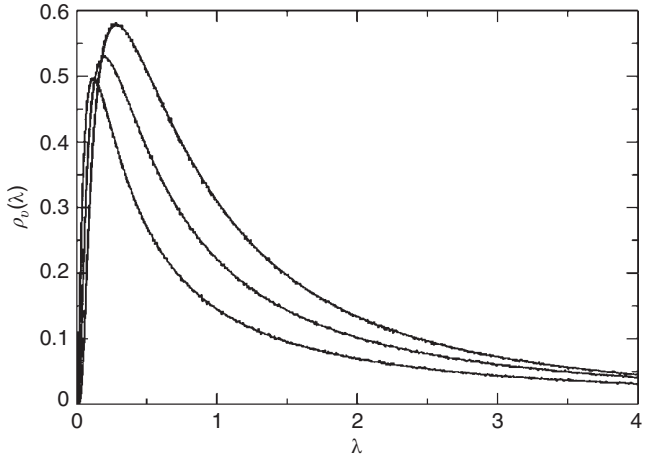


Figure 2.5 Spectra of the empirical covariance matrix **S** calculated from (2.30) with $c = 1/3$, compared to experimental data (stair lines) obtained by the Monte Carlo generation of finite matrices $N = 50, T = 150$. Source: Reproduced from [124] with permission.

for dealing with random variables that do not commute, examples of which are random matrices.

On the other hand, free random variables were initiated by Voiculescu *et al.* in 1992 and Speicher in 1994 as a rather abstract approach to von Neumann algebras, but it has a *concrete realization* in the context of RMT, because large random matrices can be regarded as free random variables, as mentioned above.

The centerpiece of free random variables is a mathematical construction of the notion of *freeness*, which is a noncommutative counterpart of the classical **independence** of random variables. As such, it allows for extending many classical results founded upon the properties of **independence** into the noncommutative (random matrix) realm, particularly the algorithms of addition and multiplication of random variables, or the ideas of stability, infinite divisibility, and so forth. This introduces a new quality into RMT, which simplifies, both conceptually and technically, many random matrix calculations, especially in the macroscopic limit (the bulk limit, or random matrices of infinite size), which is the main interest in practical problems.

We formulate an analog of the central limit theorem, if random variables $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ forming the sums

$$\mathbf{S}_n = \mathbf{X}_1 + \mathbf{X}_2 + \cdots \mathbf{X}_n \tag{2.31}$$

do not commute. In other words, we are seeking a theory of probability, which is noncommutative, i.e. $\mathbf{X}_i$, can be viewed as operators, but which should exhibit close similarities to the "classical" theory of probability. Such theories are certainly interesting from the point of view of quantum mechanics or noncommutative field theory. Abstract operators may have matrix representations. When such a construction exists, we have a natural tool for formulating the probabilistic analysis directly in the space of matrices. Contemporary financial markets are characterized by collecting and processing enormous amount of data—big data. Statistically, they may obey the matrix central limit theorems. Matrix-valued probability theory is then ideally suited for analyzing the properties of arrays of data.

The origins of noncommutative probability are linked with abstract studies of von Neumann algebras done in the 1980s. A new twist was given to the theory when it was realized that noncommuting abstract operators, called free random variables, can be represented as infinite matrices [126]. Only very recently the concept of FRV started to appear explicitly in physics [127–129].

Below we abandon a formal way and we shall follow the intuitive approach, using frequently a physical intuition. Our main goal is to study the spectral properties of large arrays of data.

Let us assume that we want to study statistical properties of infinite random matrices. We study the spectral properties of $N \times N$ matrix $\mathbf{X}$, (in the limit N), which is drawn from a matrix measure

$$d\mathbf{X} \exp\left[-N \operatorname{Tr} V(\mathbf{X})\right] \tag{2.32}$$

with a potential $V(\mathbf{X})$ (in general not necessarily polynomial). For the moment, we study real symmetric matrices whose spectrum is real. The average spectral density of the matrix $\mathbf{X}$ is defined as

$$\rho(\lambda) = \frac{1}{N} \langle \operatorname{Tr} \delta(\lambda - \mathbf{X}) \rangle = \frac{1}{N} \left\langle \sum_{i=1}^{N} \delta\left(\lambda - \lambda_i\right) \right\rangle \tag{2.33}$$

where $\langle \cdots \rangle$ means averaging over the ensemble (2.32) and $\lambda_i = \lambda_i(\mathbf{X})$ are the eigenvalues of $\mathbf{X}$. Using the standard folklore, that the spectral properties are related to the discontinuities of the Green's function we may introduce

$$G(z) = \frac{1}{N} \left\langle \operatorname{Tr} \frac{1}{z\mathbf{I} - \mathbf{X}} \right\rangle \tag{2.34}$$

where z is a complex variable and $\frac{1}{z\mathbf{I}-\mathbf{X}}$ stands for the inverse $(z\mathbf{I}-\mathbf{X})^{-1}$. Due to the known properties of the distributions

$$\lim_{\varepsilon\to 0}\frac{1}{\lambda\pm i\varepsilon}=PV\frac{1}{\lambda}\mp i\pi\delta(\lambda) \tag{2.35}$$

we find that the imaginary part of the Green's function reconstructs spectral density (2.33)

$$-\frac{1}{\pi}\lim_{\varepsilon\to 0}\operatorname{Im}G(z)\big|_{z=\lambda+i\varepsilon}=\rho(\lambda) \tag{2.36}$$

This famous inversion formula motivates the whole framework.

The natural Green's function will serve as an auxiliary construction explaining the crucial concepts of the theory of matrix (noncommutative) probability theory. Let us define a functional inverse of the Green's function (sometimes called a Blue's function [128]), i.e., $G[B(z)]=z$. The fundamental object in noncommutative probability theory, the so-called R function or R-transform, is defined as

$$R(z)=B(z)-\frac{1}{z} \tag{2.37}$$

With the help of the R-transform, we shall now uncover several astonishing analogies between the classical and matrix probability theory.

We shall start from the analog of the central limit theorem [126]: the spectral distributions of independent variables $\mathbf{X}_i$

$$\mathbf{S}_K=\frac{1}{\sqrt{K}}\left(\mathbf{X}_1+\mathbf{X}_2+\cdots\mathbf{X}_K\right) \tag{2.38}$$

each with arbitrary probability measure with zero mean and finite variance $\left\langle \operatorname{Tr}\mathbf{X}_i^2\right\rangle=\sigma^2$, converge towards the distribution with R-transform $R(z)=\sigma^2 z$.

Let us now find the exact form of this limiting distribution. Since $R(z)=\sigma^2 z$, $B(z)=\sigma^2 z+1/z$, its functional inverse satisfies

$$z=\sigma^2 G(z)+1/G(z) \tag{2.39}$$

The solution of this quadratic equation (with proper asymptotics $G(z)\to 1/z$ for large z) is

$$G(z)=\frac{z-\sqrt{z^2-4\sigma^2}}{2\sigma^2} \tag{2.40}$$

so the spectral density, supported by the cut of the square root, is

$$\rho(\lambda)=\frac{1}{2\pi\sigma^2}\sqrt{4\sigma^2-\lambda^2} \tag{2.41}$$

This is the famous Wigner semicircle [102] (actually, semiellipse) ensemble. The omnipresence of this ensemble in various physical applications finds a natural explanation—it is a consequence of the central limit theorem for noncommuting random variables. Thus the Wigner ensemble is a noncommutative analog of the Gaussian distribution. Indeed, one can show that the measure (2.32) corresponding to the Green's function (2.40) is $V(\mathbf{X})=\frac{1}{\sigma^2}\mathbf{X}^2$.

Now consider, what "independence" means for two identical matrix valued ensembles, for example, of the Gaussian type, with zero mean and unit variance. We intend to find the discontinuities of the Green's function

$$G_{1+2}(z) \sim \int D\mathbf{X}_1 D\mathbf{X}_2 e^{-N \operatorname{Tr} \mathbf{X}_1^2} e^{-N \operatorname{Tr} \mathbf{X}_2^2} \operatorname{Tr} \frac{1}{z\mathbf{I} - \left(\mathbf{X}_1 + \mathbf{X}_2\right)} \tag{2.42}$$

In principle, this requires a solution of the convolution, with matrix-valued, noncommuting entries. Here we can see how the R-transform operates. This is the transform that imposes the additive property for the all cumulants: all spectral cumulants obey

$$k_i\left(\mathbf{X}_1 + \mathbf{X}_2\right) = k_i\left(\mathbf{X}_1\right) + k_i\left(\mathbf{X}_2\right)$$

for all $i = 1, 2, \ldots, \infty$ [126, 130].

Mathematicians call such a property "freeness," hence the name "free random variables." The R-transform is an analog of the logarithm of the characteristic function in the classical probability theory, and fulfills the addition law [126]

$$R_{1+2}(z) = R_1(z) + R_2(z) \tag{2.43}$$

For two large random matrices $\mathbf{X}$, Y, we have

$$R_{\mathbf{X}+\mathbf{Y}}(z) = R_{\mathbf{X}}(z) + R_{\mathbf{Y}}(z) \tag{2.44}$$

At this moment one can start to really appreciate the power of the noncommutative approach to probability. For large random matrices $\mathbf{X}$ and $\mathbf{Y}$ (exact results hold in the $N \to \infty$ limit), the knowledge of their spectra is usually sufficient for predicting the *spectrum* of the sum $\mathbf{X} + \mathbf{Y}$.

The noncommutative calculus also allows generalization of the additive law for non-Hermitian matrices [123, 131], and even the formulation of the multiplicative law, inferring the knowledge of all moments of the spectral function of the product of $\mathbf{XY}$, knowing only the spectra of $\mathbf{X}$ and $\mathbf{Y}$ separately (so-called S-transform) [126]. It turns out that for two large random matrices $\mathbf{X}$, Y, we have

$$S_{\mathbf{XY}}(z) = S_{\mathbf{X}}(z) S_{\mathbf{Y}}(z) \tag{2.45}$$

As such, it offers a powerful shortcut in analyzing stochastic properties of large ensembles of data. Moreover, the larger the sets the better because finite size affects scale at least as $1/N$.

Consider power-law like spectra in noncommutative probability theory. Motivated by the construction in classical probability, we pose the following problem: what is the most general form of the spectral distribution of random matrix ensemble, which is stable under matrix convolution, with the same functional form as the original distributions, modulo shift, and rescaling? Surprisingly, noncommutative probability theory follows from the Lévy–Khinchine theorem of stability in classical probability. In general, the required $R(z)$ behaves like $z^{\alpha-1}$, where $\alpha \in (0, 2]$. More precisely, the list is exhausted by the following R-transforms [132]:

- (i) $R(z) = e^{i\pi\phi} z^{\alpha-1}$, where $\alpha \in (1, 2], \phi[\alpha - 2, 0]$
- (ii) $R(z) = e^{i\pi\phi} z^{\alpha-1}$, where $\alpha \in (0, 1], \phi[1, 1 + \alpha]$
- (iii) $R(z) = a + b \log z$, where b is real, $\Im a \geq 0$ and $b \geqslant -\frac{1}{\pi} \operatorname{Im} a$

The asymptotic form of the spectra is power-law like, i.e., $\rho(\lambda) \sim 1/\lambda^{\alpha-1}$. The singular case (iii) corresponds, in a symmetric case ($b = 0$), to the Cauchy distribution. Note that case (i) with $\alpha = 2$ corresponds to the Gaussian ensemble. For spectral distributions, several other analogies to Levy distributions hold. In particular, there is a one-to-one correspondence for spectral analogs of ranges, asymmetries, and shifts. Spectral distributions also exhibit duality laws ($\alpha \to 1/\alpha$), like their classical counterparts [133, 134].

Let us show how useful the formalism of noncommutative probability theory could be for the analysis of financial data.

We analyze a time series of prices of N companies, measured at equal sequence of T intervals. The returns (here relative daily changes of prices) could be recast into $N \times T$ matrix $\mathbf{X}$. This matrix defines the empirical $N \times N$ covariance matrix $\mathbf{C}$. This matrix today forms a cornerstone of every methodology of measuring the market risk.

Now we are ready to confront the empirical data. Consider the extreme case in which the covariance matrix is completely noisy (no information), i.e., $\mathbf{X}$ is stochastic, belonging to a random matrix ensemble. By central limit theorems, we can consider either matrix Gaussian or matrix Lévy–Khinchine stability basins. The exact formula, corresponding to $T, N \to \infty$, with $N/T = c$ fixed, comes from [131].

For symmetric Levy distributions, for completely random matrices, the Green's function is given by

$$G(z) = 1/z\left[1 + f(z)\right] \tag{2.46}$$

where $f(z)$ is a multivalued solution of a transcendental equation

$$\left(1 + f\right)\left(f + c\right)\frac{1}{f^{2/\alpha}} = z \tag{2.47}$$

In the case where $\alpha = 2$, (2.47) is algebraic (quadratic), and the spectrum is localized on a finite interval. In all other cases, the range of the spectrum is infinite, with the large eigenvalue distribution scaling as $1/\lambda^{\alpha+1}$.

The case in which $\alpha = 2$ corresponds to the spectral distribution of celebrated the Marchenko–Pastur law. See Section 2.10.2.

In the case of a Gaussian disorder, 94% of empirical eigenvalues were consistent with random matrix spectra, as pointed out in Section 2.10.2. Only a few of the largest eigenvalues did not match the pattern, reflecting the appearance of large clusters of companies. The analysis done with the power law ($\alpha = 1.5$) not only confirmed the dominance of stochastic effects, but even interpreted the clusters as possible large stochastic events [134]. It also pointed out the dangers of using the covariance matrix (which assumes implicitly the finite dispersion) in cases when power laws are present. A comparative study shows that only this covariance is stable under reshuffling, with a spectrum in remarkable agreement with the one extracted from an ensemble of random Lévy matrices with commensurate sizes and asymmetry.

2.10.8 Cross-Correlations between Input and Output Variables

Our central result is derived from the theory of free random matrices, and gives an explicit expression for the interval where singular values are expected in the absence of any true correlations between the variables under study. Our result can be seen as the natural generalization of the Marchenko–Pastur distribution for the case of rectangular correlation matrices.

Consider the N input factors, denoted as $X_i, i = 1, \ldots, N$ and M output factors $Y_j, j = 1, \ldots, M$. There are a total of T observations, where X_{it} and $Y_{jt}, t = 1, \ldots, T$ are observed. We assume that all $N + M$ time series are standardized—that both Xs and Ys have zero mean and variance unity. The Xs and the Ys may be completely different, or may be the same set of observables but observed at different times, as for example $N = M$ and $Y_{jt} = X_{it+1}$.

Now, consider the $M \times N$ cross-correlation matrix $\mathbf{R}$ between the Xs and the Ys:

$$(\mathbf{R})_{ij} = \sum_{t=1}^{T} Y_{jt} X_{it} \equiv \left(\mathbf{Y}\mathbf{X}^T\right)_{ij} \tag{2.48}$$

We are interested in the singular value decomposition (SVD) of this matrix. If $M < N$, we consider the matrix $M \times M$ matrix $\mathbf{R}\mathbf{R}^T$ (or the $N \times N$ matrix $\mathbf{R}^T\mathbf{R}$ if $M > N$), which is symmetrical and has M positive eigenvalues, each of which is equal to the square of a singular value of $\mathbf{R}$ itself. The second observation is that the nonzero eigenvalues of $\mathbf{R}\mathbf{R}^T = \mathbf{Y}\mathbf{X}^T\mathbf{X}\mathbf{Y}^T$ are the same as those of the $T \times T$ matrix $\mathbf{T} = \mathbf{X}^T\mathbf{X}\mathbf{Y}\mathbf{Y}^T$, obtained by swapping the position of $\mathbf{Y}$ from the first to the last. In the benchmark situation (null hypothesis) where the Xs and the Ys are independent from each other, the two matrices $\mathbf{X}^T\mathbf{X}$ and $\mathbf{Y}\mathbf{Y}^T$ are mutually free [52], and one can use results on the product of free matrices to obtain the eigenvalue density from that of the two individual matrices, which are known. The general recipe [52, 135] is to construct first the so-called η-transform of the eigenvalue density $\rho(u)$ of a given $T \times T$ non-negative matrix $\mathbf{A}$, defined as:

$$\eta(\gamma) = \int du \frac{1}{1 + \gamma u} \equiv \frac{1}{T} \operatorname{Tr}\left(\mathbf{I}_T + \gamma \mathbf{A}\right)^{-1} \tag{2.49}$$

From the functional inverse of $\eta(\gamma)$, one now defines the S-transform of $\mathbf{A}$ as:

$$S_{\mathbf{A}}(x) \equiv -\frac{1 + x}{x} \eta_{\mathbf{A}}^{-1}(1 + x) \tag{2.50}$$

Endowed with these definitions, one of the fundamental theorems of free matrix theory [52] states that the S-transform of the product of two free matrices $\mathbf{A}$ and $\mathbf{B}$ is equal to the product of the two S-transforms

$$S_{\mathbf{AB}}(x) = S_{\mathbf{A}}(x) \Sigma_{\mathbf{B}}(x)$$

A similar, somewhat simpler, theorem exists for sums of free matrices, in terms of R-transforms, such that

$$R_{\mathbf{A+B}}(x) = R_{\mathbf{A}}(x) + R_{\mathbf{B}}(x)$$

Applying this theorem with $\mathbf{A} = \mathbf{X}^T\mathbf{X}$ and $\mathbf{B} = \mathbf{Y}\mathbf{Y}^T$ one finds:

$$\begin{aligned} \eta_{\mathbf{A}}(\gamma) &= 1 - n + \frac{n}{1 + \gamma}, \quad n = \frac{N}{T} \\ \eta_{\mathbf{B}}(\gamma) &= 1 - m + \frac{m}{1 + \gamma}, \quad m = \frac{M}{T} \end{aligned} \tag{2.51}$$

From this, one easily obtains:

$$S_{\mathbf{T}}(x) = S_{\mathbf{X}^T\mathbf{X}\mathbf{Y}\mathbf{Y}^T}(x) = S_{\mathbf{X}^T\mathbf{X}}(x) S_{\mathbf{Y}\mathbf{Y}^T}(x) = \frac{(1 + x)^2}{(x + n)(x + m)} \tag{2.52}$$

Inverting back this relation allows one to derive the η-transform of $\mathbf{T}$ as:

$$\eta_{\mathbf{T}}(\gamma) = \frac{1}{2(1+\gamma)}\left[1-(\mu+\nu)\gamma+\sqrt{(\mu-\nu)^2\gamma^2-2(\mu+\nu+2\mu\nu)\gamma+1}\right] \quad (2.53)$$

with $\mu = m-1$ and $\nu = n-1$. The limit $\gamma \to \infty$ of this quantity gives the density of exactly zero eigenvalues, easily found to be equal to $\max(1n, 1m)$, meaning, as expected, that the number of nonzero eigenvalues of $\mathbf{T}$ is $\min(N, M)$. Depending on the value of $n+m$ compared to unity, the pole at $\gamma = 1$ corresponding to eigenvalues exactly equal to 1 has a zero weight (for $n+m<1$) or a nonzero weight equal to $n+m$. One can rewrite the above result in terms of the more common Stieltjes transform of $\mathbf{T}$, $m_{\mathbf{A}}(z) \equiv \eta_{\mathbf{A}}(-1/z)/z$, for matrix $\mathbf{A}$, which reads:

$$m_{\mathbf{T}}(z) = \frac{1}{2z(z-1)}\left[z+(\mu+\nu)+\sqrt{(\mu-\nu)^2-2(\mu+\nu+2\mu\nu)z+z^2}\right] \quad (2.54)$$

The density of eigenvalues is then obtained from the standard relation [52]:

$$\rho_{\mathbf{T}}(z) = \lim_{\varepsilon\to 0}\mathrm{Im}\left[\frac{1}{\pi T}\mathrm{Tr}\left((z+i\varepsilon)\mathbf{I}_T-\mathbf{T}\right)^{-1}\right] = \lim_{\varepsilon\to 0}\frac{1}{\pi}\mathrm{Im}\left[m_{\mathbf{T}}(z)\right] \quad (2.55)$$

which leads to the rather simple final expression, which is the central result of this section, for the density of singular values s of the original correlation matrix $\mathbf{R} = \mathbf{Y}\mathbf{X}^T$:

$$\rho(s) = \max(1-n, 1-m)\,\delta(s) + \max(m+n-1, 0)\,\delta(s-1) + \frac{\mathrm{Re}\sqrt{(s^2-\gamma_-)(\gamma_+-s^2)}}{\pi s(1-s^2)} \quad (2.56)$$

where $\gamma_\pm$ are the two positive roots of the quadratic expression under the square root in equation (2.54) above, which read explicitly:

$$\gamma_\pm = n+m-2mn \pm 2\sqrt{mn(1-n)(1-m)} \quad (2.57)$$

This is our main technical result.

We can choose as a benchmark the case where all (standardized) variables X and Y are uncorrelated, meaning that the ensemble average $\mathbb{E}(\mathbf{C}_X) = \mathbb{E}(\mathbf{X}\mathbf{X}^T)$ and $\mathbb{E}(\mathbf{C}_Y) = \mathbb{E}(\mathbf{Y}\mathbf{Y}^T)$ are equal to the unit matrix, whereas the ensemble average cross-correlation $\mathbb{E}(\mathbf{R}) = \mathbb{E}(\mathbf{X}\mathbf{Y}^T)$ is identically zero.

For a given finite size sample, however, the eigenvalues of $\mathbf{C}_X$ and $\mathbf{C}_Y$ will differ from the limiting value (unit), and the singular values of $\mathbf{R}$ will not be the limiting value (zero). The statistics of the eigenvalues of $\mathbf{C}_X$ and $\mathbf{C}_Y$ is well known to be given by the Marchenko–Pastur distribution with parameters n and m respectively, which reads, for $c = n, m < 1$:

$$\rho_{MP}(\lambda) = \frac{1}{2\pi c\lambda}\mathrm{Re}\sqrt{(\lambda-\lambda_{\min})(\lambda_{\max}-\lambda)} \quad (2.58)$$

with

$$\lambda_{\min} = \left(1-\sqrt{c}\right)^2 \quad \lambda_{\max} = \left(1+\sqrt{c}\right)^2 \quad (2.59)$$

The S-transform of this density takes a particularly simple form:

$$S(x) = \frac{1}{1+cx} \quad (2.60)$$

The singular values of $\mathbf{R}$ are obtained as the square-root of the eigenvalues of $\mathbf{T} = \mathbf{X}^T\mathbf{X}\mathbf{Y}\mathbf{Y}^T$. Since $\mathbf{X}^T\mathbf{X}$ and $\mathbf{Y}^T\mathbf{Y}$ are mutually free, one can again use the multiplication rule of S-transforms, after having noted that the S-transform of the $T \times T$ matrices $\mathbf{X}^T\mathbf{X}$ and $\mathbf{Y}^T\mathbf{Y}$ is now given by:

$$S(x) = \frac{1}{x+c} \tag{2.61}$$

One therefore finds that the η-transform of $\mathbf{T}$ is obtained by solving the following cubic equation for x:

$$\eta^{-1}(1+x) = -\frac{1+x}{x(n+x)(m+x)} \tag{2.62}$$

which can be done explicitly, leading to the following (lengthy) result. To denote $y = s^2$, one should first compute the following two functions:

$$f_1(y) = 1 + m^2 + n^2 - mn - m - n + 3y \tag{2.63}$$

and

$$\begin{aligned} f_2(y) = {} & 2 - 3m(1-m) - 3n(1-n) - 3mn(n+m-4) \\ & + 2\left(m^3 + n^3\right) + 9y(1+m+n) \end{aligned} \tag{2.64}$$

Then, form

$$\Delta = -4f_1(y)^3 + f_2(y)^2 \tag{2.65}$$

If $\Delta > 0$, one introduces a second auxiliary variable Γ:

$$\Gamma = f_2(y) - \sqrt{\Delta} \tag{2.66}$$

to compute $\rho_2(y)$:

$$\pi\rho_2(y) = -\frac{\Gamma^{1/3}}{2^{4/3}3^{1/2}y} + \frac{f_1(y)}{2^{2/3}3^{1/2}\Gamma^{1/3}y} \tag{2.67}$$

Finally, the density $\rho(s)$ is given by:

$$\rho(s) = 2s\rho_2\left(s^2\right) \tag{2.68}$$

2.11 Big Data for Atmospheric Systems

Here we show that the empirical correlation matrices that arise in atmospheric sciences can be modeled as a random matrix chosen from an appropriate ensemble. The correlation studies are elegantly carried out in the matrix framework.

Weather and climate data are frequently subjected to principal component analysis (via singular value decomposition) to identify the independent modes of atmospheric variability. The analysis performed on the correlation matrices is aimed at separating the signal from "noise," to cull the physically meaningful modes of the correlation matrix from the underlying noise.

The empirical orthogonal function (EOF) method, also called principal component analysis, is a multivariate statistical technique widely used in the analysis of geophysical data. It is similar to the singular value decomposition employed in linear algebra and

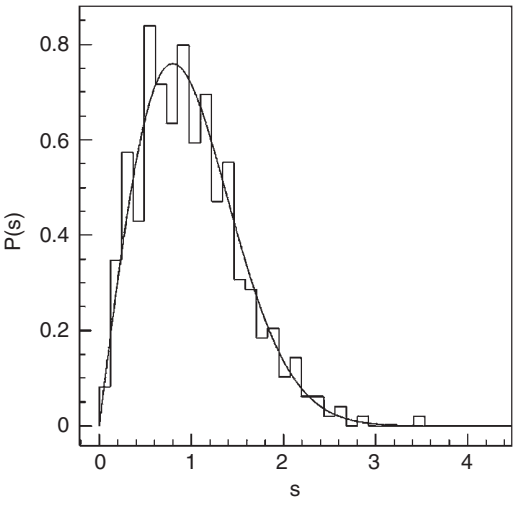


Figure 2.6 Eigenvalue spacing distribution for the monthly mean sea-level pressure (SLP) correlation matrix. The solid curve is the GOE prediction. Source: Reproduced with permission from [106].

it provides information about the independent modes of variabilities exhibited by the system.

In general, any atmospheric parameter $z(x, t)$ (like wind velocity, geopotential height, temperature, etc.) varies with space x and time t and is assumed to follow an average trend on which the variations (or anomalies as referred to in atmospheric sciences) are superimposed: $z(x, t) = z_{avg}(x) + z'(x, t)$.

If the observations were taken n times at each of the p spatial locations and the corresponding anomalies $z'(x, t)$ assembled in the data matrix $\mathbf{Z}$ of order $p \times n$, then the spatial correlation matrix of the anomalies is given by

$$\mathbf{S} = \frac{1}{n}\mathbf{Z}\mathbf{Z}^H \tag{2.69}$$

The elements of the Hermitian matrix $\mathbf{S}$ of order p are just the Pearson correlation between various spatial points. The eigenfunctions of $\mathbf{S}$ are called the empirical orthogonal functions because they form a complete set of orthogonal basis to represent the data matrix $\mathbf{Z}$. The size of $\mathbf{Z}$ is large. For example $n = 500, p = 600$.

We will show that the spectrum of $\mathbf{S}$ displays random-matrix-type spectral statistics. In [106] Santhanam and Patra have analyzed atmospheric correlation matrices from the perspective of the random matrix theory. The central result of their work is that atmospheric correlation matrices can be modeled as random matrices chosen from an appropriate RMT ensemble. The spectrum of atmospheric correlation matrices satisfy the random matrix prescription, as shown in Figure 2.6 and 2.7. In particular, the eigenmodes of the atmospheric empirical correlation matrices that have physical significance are marked by deviations from the eigenvector distribution.

2.12 Big Data for Sensing Networks

Our vision for big data follows Figure 1.6, first suggested in [39, 40]. The mathematical foundation of big data is treated in [40], with sensing networks being the motivated application.

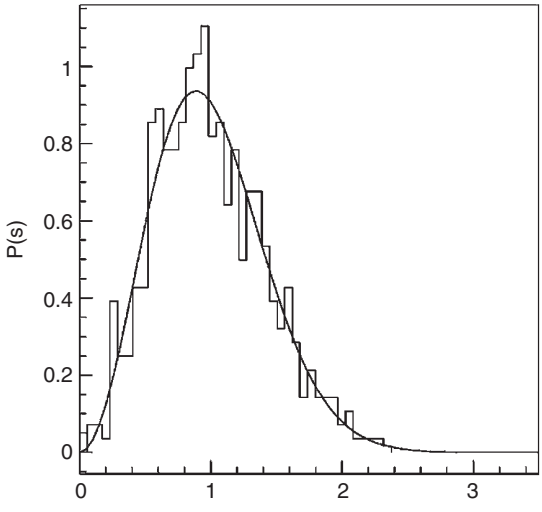


Figure 2.7 Eigenvalue spacing distribution for the monthly mean wind-stress correlation matrix. The solid curve is the GUE prediction. Source: Reproduced with permission from [106].

2.13 Big Data for Wireless Networks

Our vision for big data follows Figure 1.6, first suggested in [39, 40]. The treatment of cognitive radio networks as a big data problem is addressed in [39]. Here we highlight some applications of this new methodology [44]. See [70] for other applications.

In the spirit of our previous work [40]—representing large datasets in terms of random matrices—we report some empirical findings here. In this initial report, we summarize the most interesting results only when the theoretical models agree with experimental data. When the size of a random matrix is sufficiently large, the empirical distribution of the eigenvalues (viewed as functions of this random matrix) converges to some theoretical limits (such as Marchenko-Pastur law and the single ring law). In the context of large-scale wireless network, our empirical findings will validate these theoretical predictions. To the best of our knowledge, our work represents the first such attempt in the literature, although a lot of simulations have been used in earlier work [136].

2.13.1 Marchenko–Pastur Law

Let $\mathbf{X} = \left\{\xi_{ij}\right\}_{1\leqslant i\leqslant N, 1\leqslant j\leqslant n}$ be a random $N \times n$ matrix whose entries are i.i.d. N is an integer such that $N \leqslant n$ and $N/n = c$ for some $c \in (0, 1]$. The empirical spectrum density (ESD) of the corresponding sample covariance matrix $\mathbf{S} = \frac{1}{n}\mathbf{X}^H\mathbf{X}$ converges to the distribution of the Marchenko–Pastur law [39, 136] with density function

$$f_{MP}(x) = \begin{cases} \frac{1}{2\pi x c\sigma^2}\sqrt{(b-x)(x-a)}, & a \leqslant x \leqslant b \\ 0 & otherwise \end{cases} \tag{2.70}$$

where $a = \sigma^2(1-\sqrt{c})^2, \quad b = \sigma^2(1+\sqrt{c})^2$

2.13.2 The Single "Ring" Law

For each $n \geq 1$, let $\mathbf{A}_n$ be a random matrix that admits the decomposition $\mathbf{A}_n = \mathbf{U}_n \mathbf{T}_n \mathbf{V}_n$, with $\mathbf{T}_n = \text{diag}\left(s_1, \ldots, s_n\right)$ where the s_is are positive numbers and where $\mathbf{U}_n$ and $\mathbf{V}_n$ are two independent random unitary matrices, which are Haar-distributed independently from the matrix $\mathbf{T}_n$. Under certain mild conditions, the ESD $\mu_{\mathbf{A}_n}$ of $\mathbf{A}_n$ converges [137], in probability, weakly to a deterministic measure whose support is $\{z \in \mathbb{C} : a \leqslant |z| \leqslant b\}$, $a = \left(\int x^{-2} v\,(dx)\right)^{-1/2}$, $b = \left(\int x^2 v\,(dx)\right)^{1/2}$. Some outliers to the single ring law [138] can be observed.

Consider the matrix product $\prod_{i=1}^{\alpha} \mathbf{X}_i$, where $\mathbf{X}_i$ is the singular value equivalent [139] of the rectangular $N \times n$ non-Hermitian random matrix $\tilde{\mathbf{X}}_i$, whose entries are i.i.d. Thus, the empirical eigenvalue distribution of $\prod_{i=1}^{\alpha} \mathbf{X}_i$ are almost certain to converge to the same limit given by

$$f_{\prod_{i=1}^{\alpha} \mathbf{X}_i}(z) = \begin{cases} \frac{1}{\pi c \alpha} |z|^{2/\alpha - 2} & (1-c)^{\alpha/2} \leqslant |z| \leqslant 1 \\ 0 & \textit{elsewhere} \end{cases} \tag{2.71}$$

as $N, n \to \infty$ with the ratio $c = N/n \leqslant 1$. On the complex plane of the eigenvalues, the inner circle radius is $(1-c)^{\alpha/2}$ and the outer circle radius is unity.

2.13.3 Experimental Results

All the data are collected under two scenarios: (i) only noise is present; or (ii) signal plus noise are present. We have used 70 USRP front ends and 29 high-performance PCs. The experiments are divided into two main categories: (i) with a single USRP receiver, and (ii) with multiple USPR receivers.

Every such software-defined radio (SDR) platform (also called a node) is composed of one or several USRP RF front ends and a high-performance PC. The RF up-conversion and down-conversion functionalities reside in the USRP front end, whereas the PC is mainly responsible for baseband signal processing. The USRP front end can be configured as either a radio receiver or a transmitter, which is connected to a PC via an Ethernet cable.

Seventy USRP receivers are organized as a distributed sensing network. One PC takes the role of the control node, which is responsible for sending a command to all the USRP receivers that will start the sensing at the same time. The network time is synchronized by the GPS attached to every USRP. The 70 USRPs are placed in random locations within a room. For every single USRP receiver, a random matrix is obtained and denoted as $\mathbf{X}_i \in \mathbb{C}^{N \times n}$, whose entries are normalized as mentioned above. We will investigate the ESD of the sum and the product of the α random matrices below, where α is the number of the random matrices.

The product of α non-Hermitian random matrices is defined as $\mathbf{Z} = \prod_{i=1}^{\alpha} \mathbf{X}_i$, where $\mathbf{X}_i \in \mathbb{C}^{N \times n}$, $i = 1, \ldots, \alpha$. A singular value equivalent is performed before multiplying the original random matrices. We are actually analyzing the empirical eigenvalue distribution as (2.71).

By specifying $\alpha = 5$, we performed the experiments for two scenarios: (i) Pure noise. In this case, neither the commercial signal nor the USRP signal is received at all the USRP receivers. Figure 2.8 shows the the ring law distribution of the eigenvalues for

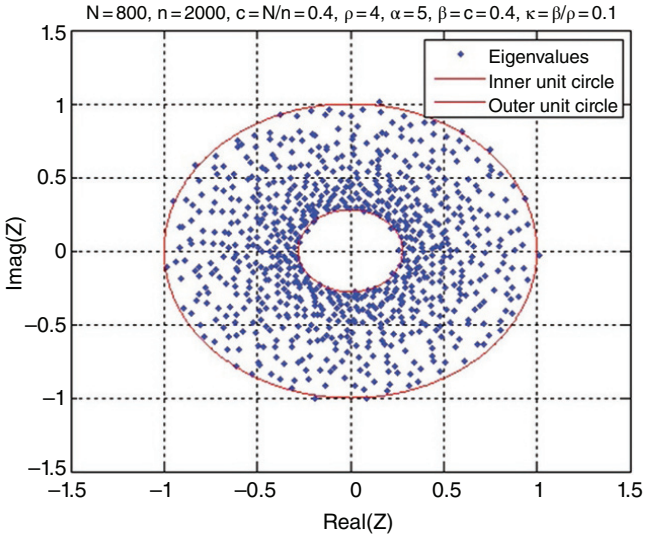


Figure 2.8 The ring law for the product of non-Hermitian random matrix with white noise only. The number of random matrix $\alpha = 5$. The radii of the inner circle and the outer circle agree with (2.71).

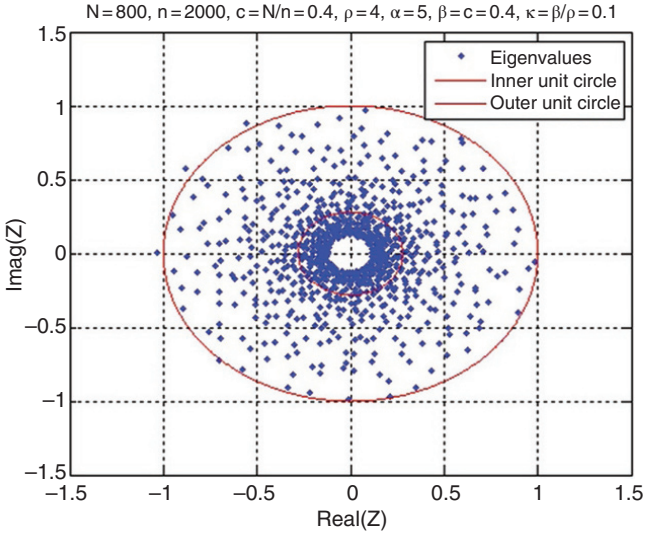


Figure 2.9 The ring law for the product of non-Hermitian random matrices with signal plus white noise. The number of random matrix $\alpha = 5$. The radius of the inner circle is less than that of the white-noise-only scenario.

the product of non-Hermitian random matrices when α is 5. The radii of the inner circle and outer circle are well matched with the result in (2.71). (ii) Commercial signal with frequency. In this case, the signal at the frequency of 869.5 MHz is used. Figure 2.9 shows the ring-law distribution of the eigenvalues for the product of non-Hermitian

random matrices, when signal plus white noise is present. By comparing Figure 2.9 with Figure 2.8, we find that, in the signal-plus-white-noise case, the inner radius is smaller than that of the white-noise-only case.

2.14 Big Data for Transportation

In a 5G wireless communication system [140, 141], low-latency data becomes important. Vehicle-to-vehicle communications with low-latency will enable big data for transportation.

One might be interested in the following ideas: (i) aggregate data from a large number of vehicles; (ii) the kinds of statistical laws these data will follow; (iii) how to model these data using large random matrices.

Bibliographical Remarks

In Section 2.1, we drew on material from [83].
In Section 2.2, we drew on material from [142, 143, 143, 144].
In Section 2.5, we drew on material from [145], [18] and [146].
We drew on material from [106] in Section 2.9.
We followed [106] in Section 2.11.
We followed [115] in Section 2.10.4.
We followed [82, 125, 147] in Section 2.10.7.
In Section 2.10, we followed references [87, 106, 107, 115, 115, 117, 118, 125, 147–153].
We followed [119, 120, 154, 155] in Section 2.10.5.
We followed [120, 121, 124, 153, 156] for our development of Section 2.10.6.
We followed [82] in Section 2.10.8.
In Section 2.8, we drew material from [157]. Section 2.8.1 drew material from [158].
In Section 2.7, we drew material from [84, 159].

3

Large Random Matrices: An Introduction

This chapter is the basic material for next-generation engineers and researchers. It is our belief that large dimensional random matrices are the foundation for the analysis of big data; Sections 1.3 and 1.4 support this belief. We give the fundamentals of large dimensional random matrices. One motivation is to model large data sets using large random matrices. Recently there has been increasing interest in studying large dimensional data sets that arise in finance, wireless communications, genetics and other fields. Patterns in these data can often be summarized by the sample covariance matrix, as is done in multivariate regression and dimension reduction via factor analysis. In Chapter 8, for example, we apply large random matrices for anomaly detection.

Large random matrices serve as a finite-dimensional approximation of infinite-dimensional operators. Its importance for statistics comes from the fact that RMT may be used to correct traditional tests or estimators, which fail in the "large n, large p" setting, where p is the number of parameters (dimension) and n is the sample size. For example, we can use RMT for corrections on some likelihood ratio tests that fail even for moderate p (around 20) [160].

Statistical science is an empirical science. The object of statistical methods, according to R. A. Fisher (1922) [68], is the reduction of data: "It is the object of the statistical processes employed in the reduction of data to exclude this irrelevant information, and isolate the whole of the relevant information contained in the data." In the age of big data [40], the goal set by Fisher has never been so relevant as it is today. In [39], we make an *explicit* connection between big data and large random matrices. This connection is based on the simple observation that a massive amount of data can be naturally represented by (large) random matrices. When the dimensions of the random matrices are sufficiently large, some unique phenomena (such as concentration of spectral measures) will occur [40].

3.1 Modeling of Large Dimensional Data as Random Matrices

In multivariate statistics, we observe a random sample of p-dimensional observations $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ or $\mathbb{C}^p$. The statistical methods, such as principal components analysis, were developed in the early 1900s. Most results consider an asymptotic framework, where the number of observations n grows to infinity.

Smart Grid using Big Data Analytics: A Random Matrix Theory Approach, First Edition.
Robert C. Qiu and Paul Antonik.

Most of these results assume that the dimension p of the variables is fixed and "small" (less than 10 generally), whereas the number of observations n tends to infinity, $n \to \infty$. This is the classical asymptotic theory. The coming of big data mandates the analysis of high-dimensional data. The dimension p of the data is quite far away from classical situations where p is lower than 10. This new type of data is called "large dimensional data." The most remarkable fact is that n and p are large and comparable. One wonders what happens if one considers the asymptotic regime

$$n \to \infty, p \to \infty, \text{ but } \frac{p}{n} \to c \in (0, \infty) \tag{3.1}$$

So-called random matrix theory is the natural answer to this question.

Let us use an example to illustrate this point. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a Gaussian sample $\mathcal{N}\left(\mathbf{0}, \mathbf{I}_p\right)$ of dimension p, with zero mean and identity covariance matrix (also called population covariance matrix). The associated sample covariance matrix $\mathbf{S}_n$ is defined by

$$\mathbf{S}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^H$$

An important statistic in multivariate analysis is

$$T_n = \log\left(\det \mathbf{S}_n\right) = \sum_{i=1}^{p} \log \lambda_{n,i}$$

where $\left\{\lambda_{n,j}\right\}_{1 \leqslant j \leqslant p}$ are the eigenvalues of $\mathbf{S}_n$. If p is kept fixed, then $\lambda_{n,i} \to 1$ almost surely as $n \to \infty$ and thus $T_n \to 0$. Besides, by taking a Taylor expansion of $\log(1 + x)$, one can show that, for any p fixed

$$\sqrt{\frac{n}{p}} T_n \xrightarrow{\mathcal{L}} \mathcal{N}(0, 2)$$

This suggests the possibility that T_n remains asymptotically normal for a large p, assuming that $p = O(n)$. However, this is not the case: if we assume that $p/n \to c \in (0, 1)$, as $n \to \infty$, using results on an empirical spectral distribution of $\mathbf{S}_n$ (see Example 3.5.2.). It can be proved that, almost certainly

$$\sqrt{\frac{1}{p}} T_n \to \int_a^b \frac{\log x}{2\pi c x} \sqrt{(b - x)(x - a)} dx = \frac{c - 1}{c} \log(1 - c) = d(c) < 0$$

where $a = \left(1 - \sqrt{c}\right)^2$ and $b = \left(1 + \sqrt{c}\right)^2$. Thus, almost certainly

$$\sqrt{\frac{n}{p}} T_n \simeq d(c) \sqrt{np} \to -\infty$$

Consequently, any test that assumes an asymptotic normality of T_n will lead to a serious error.

This example shows that the classical large sample limits are no longer suitable for dealing with large dimensional data. Statisticians must seek out new limiting theorems instead. Thus, the theory of random matrices (RMT) might be one possible method for this aim.

3.2 A Brief of Random Matrix Theory

What do the eigenvalues of a typical large matrix look like? Do we expect certain universal patterns of eigenvalue statistics to emerge? Large complex systems often exhibit remarkably simple universal patterns as the number of degrees of freedom increases. The simplest example is the central limit theorem: the fluctuation of the sums of independent random (scalar-valued) variables, irrespective of their distributions, follows the Gaussian distribution. The other cornerstone of probability theory identifies the Poisson point process as the universal limit of many independent point-like events in space or time. These mathematical descriptions assume that the original system has independent (or at least weakly dependent) constituents. What if independence is not a realistic approximation and strong correlations need to be modelled? Is there a universality for strongly correlated models?

At first sight this seems an impossible task. While independence is a unique concept, correlations come in many different forms; there is no reason to believe that they all behave similarly. Nevertheless they do. The actual correlated system Wigner studied was the energy levels of heavy nuclei. He asked a question: what about the distribution of the rescaled energy gaps? He discovered that the difference in consecutive energy levels, after rescaling with the local density, shows a surprisingly *universal* behavior.

Wigner not only predicted universality in complicated systems but he also discovered a remarkably simple mathematical model for this new phenomenon: the eigenvalues of large random matrices. For practical purposes, infinite-dimensional Hamilton operators of quantum models are often approximated by large but finite matrices that are obtained from some type of discretization of the original continuous model. These matrices have specific forms dictated by physical rules. Without the precise knowledge of the Hamiltonian in question, the problem is still extremely difficult. Fortunately, if one is willing to lower the bar to understanding statistical properties of the eigenvalues then one can make statistical assumptions on the Hamiltonian, as long as it is consistent with the symmetry observed by the Hamiltonian. The most basic symmetry assumption one can make is that the statistical properties should be invariant under the unitary group. In more common language it means that the properties of the atom should be the same regardless of the (arbitrary) coordinate system one chooses.

Wigner's idea is far reaching. For centuries, the primary territory of probability theory was to model uncorrelated or weakly correlated systems. The surprising ubiquity of random matrix statistics is a strong evidence that it plays a similar fundamental role for correlated systems as the Gaussian distribution and the Poisson point process play for uncorrelated systems. Random matrix theory seems to provide essentially the *only universal and generally computable pattern for complicated correlated systems*. It is based on this observation that we reach our belief in RMT, upon which many parts of the whole building of big data can be grounded, to understand **correlations** of the large, complex big data system. Eugene Wigner's revolutionary vision predicted that the energy levels of large complex quantum systems exhibit a universal behavior: the statistics of energy gaps depend only on the basic symmetry type of the model. These universal statistics show *strong correlations* in the form of level repulsion and they seem to represent a new paradigm of point processes that are characteristically different from the Poisson statistics of independent points.

Random matrices have been intensively studied since the mid-1990s. In the early 1980s, major contributions on the existence of limiting spectral distributions and their explicit forms for certain classes of random matrices were made. In recent years, research on random matrix theory is turning toward second-order limiting theorems, such as the central limit theorem for linear spectral statistics, the limiting distributions of spectral spacings, and extreme eigenvalues (thus outliers). Random matrices were introduced by Wishart [116] in 1928 in mathematical statistics and started to gain more momentum after Wigner [109, 110, 161, 162]. For many years, the standard text for random matrix theory was [103], whose first edition was printed in 1967. Recently, we have seen several excellent books [35, 67, 163]. In particular, we have seen applications of RMT in wireless communication [39, 52, 136] and sensing [40].

According to quantum mechanics, the energy levels of a system are supposed to be described by the eigenvalues of a Hermitian operator H, called the Hamiltonian. To avoid the difficulty of working with an infinite-dimensional Hilbert space, we approximate the true Hilbert space by one having a finite, though large, number of dimensions.

From the very beginning, we shall make statistical hypotheses with H. Choosing a complete set of functions as basis, we represent the Hamiltonian operators H as matrices. The elements of these matrices are random variables whose distributions are restricted only by the general symmetrical properties we might impose on the ensemble of operators [103]. The problem is to obtain information on the behavior of its eigenvalues.

Consider the big-data measurement system that collects massive data sets. Here we make an analogy between the quantum system and the big-data measurement system (see Table 3.1). Suppose the massive data sets are described by an infinite-dimensional operator G. Any system must satisfy quantum mechanics; thus the massive data sets must also satisfy quantum mechanics. Our idea is to use the analogy to replace the infinite-dimensional operators G with large, but finite, dimensional random matrices $\mathbf{X}$.

In general, we use the parameter β to denote the number of standard real normals and thus $\beta = 1; 2; 4$ correspond to real, complex and quaternion respectively. $G_\beta(m, n)$ can be generated by the MATLAB command shown in Table 3.2. If $\mathbf{A}$ is an $m \times n$ random matrix $G_\beta(m, n)$ then its joint element density is given by

$$\frac{1}{(2\pi)^{\beta mn/2}} \exp\left(-\frac{1}{2}\|\mathbf{A}\|_F\right)$$

where $||\cdot||_F$ is the Frobenius norm of a matrix.

For the semicircle law, clever computational tricks for all eigenvalues [164] allow the space of $\mathcal{O}(n)$ and computation time $\mathcal{O}(n^2)$, rather than the space of $\mathcal{O}(n^2)$ and computation time $\mathcal{O}(n^3)$ of the naive computation (in MATLAB):

`A=randn(n,n); v=eig((A+A')/sqrt(2*n)).` Algorithms developed in [164] allow one to compute the largest eigenvalues of a billion by billion matrix.

The most well studied random matrices have names such as Gaussian, Wishart, MONOVA, and circular. We prefer Hermite, Laguerre, Jacobi, and perhaps Fourier. The Hermite and Laguerre ensembles are summarized in Table 3.3.

One of the most common tools in statistical analysis is principal component analysis (PCA), which identifies the top eigenvalues and top eigenvectors of a matrix of data. What one would like to know, after performing PCA, is whether the top eigenvalue has significance or whether it is purely from randomness. Tracy and Widom [165] have

Table 3.1 An analogy between the quantum system and the big data measurement system.

Quantum system	Big data measurement system	Large (but finite) random matrices
Hamiltonian operator H	Some unknown operator G	Random matrices $\mathbf{X}$
Infinite dimensions	Infinite dimensions	Finite, though large, dimensions
A continuum and a large number of discrete levels of energy	Empirical spectrum	Discrete eigenvalues

Table 3.2 Generating the Gaussian random matrix $G_\beta(m, n)$.

β	MATLAB command
1	`G=randn (m,n)`
2	`G=randn (m,n)+j*randn (m,n)`
4	`X=randn (m,n)+j*randn (m,n);`
	`Y=randn (m,n)+j*randn (m,n);G=[X Y; -conj(Y) conj(X)]`

Table 3.3 Hermite and Laguerre ensembles.

Ensemble	Matrices	Weight function	Equilibrium measure	Numeric	MATLAB
Hermite	Wigner	$e^{-x^2/2}$	semi-circle	`eig`	`g=G(n,n);H=(g+g')/2;`
Laguerre	Wishart	$x^{\nu/2-1}e^{-x/2}$	Marcenko-Pastur	`svd`	`g=G(m,n); L=(g'*g)/m;`

shown that the top eigenvalues, properly rescaled and under reasonable assumptions that the matrix entries are sufficiently independent, follow the so-called Tracy–Widom distribution. As a result, one way to understand the question above is to do hypothesis testing against, instead of the Gaussian distribution (hypothesis $\mathcal{H}_0$), the Tracy–Widom distribution (hypothesis $\mathcal{H}_1$). Some examples for hypothesis testing using RMT are [166] and two recent books [39, 40].

The use of large random matrices for big data was explicitly proposed by the current author in November 2011 when he was writing [39]. It is known that PCA is closely related to dimension reduction, which is ubiquitous for big data applications; hypothesis testing for massive data sets is the author's long-term goal. We will formulate the hypothesis test of two alternative random matrices in Section 8.2.

If we could summarize the objects of interest in random matrix theory in one sentence, it would be to study the statistical properties of functions of matrices with random entries. Problems that are of central interest are listed here:

- *Macroscopic eigenvalue distribution.* Given a square $n \times n$ random matrix $\mathbf{A}$, let $\lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_n$ denote the singular values of $\mathbf{A}$. When the matrix is Hermitian, these are

exactly its eigenvalues. What are the properties of the normalized measure induced by $\{\lambda_i\}$? In other words, what can we say about the measure

$$\frac{1}{n}\sum_{i=1}^{n}\delta_{\lambda_i}(x)? \tag{3.2}$$

For example, is the measure compactly supported?

In probability theory, we are also interested in the $n \to \infty$ limit: For what class of random matrix does the weak limit of the empirical singular value distribution (3.2) exist? If it exists, what is the limit?

- *Mesoscopic eigenvalue distribution.* Let f be a bounded function, and I an interval on the real axis, then

$$\frac{1}{n}\sum_{\lambda_i\in I} f\left(\lambda_i\right) \to \int_I f(x)\rho(x)dx$$

either probably or almost certainly. Assume now that ρ has compact support. Let $E \in \text{supp}\,\rho, f$ a continuous bounded function and $I = \left[E - \frac{1}{n^\alpha}, E + \frac{1}{n^\alpha}\right]$ for some $0 \le \alpha < 1$, is it true that

$$\frac{1}{n\,|I|}\left|\sum_{\lambda_i\in I} f\left(\lambda_i\right) - n\int_I f(x)\rho(x)dx\right| \to 0?$$

- *Microscopic eigenvalue distribution.* Assume the eigenvalues of a random matrix lie on a compact interval I. As there are n eigenvalues in the interval, the average spacing of the eigenvalues is of the order $1/n$. Let $p_n\left(\lambda_1, \lambda_2, \ldots, \lambda_n\right)$ be the joint distribution of the eigenvalues, and let

$$p_n^{(k)}\left(\lambda_1, \lambda_2, \ldots, \lambda_k\right) = \int_{\mathbb{R}^{n-k}} p_n\left(\lambda_1, \lambda_2, \ldots, \lambda_n\right) d\lambda_{k+1}\cdots d\lambda_n$$

be the k-point correlation function. Does the k-point correlation function converge to a certain limit in the local scale? In other words for $E \in I$, does the limit

$$\lim_{n\to\infty} p_n^{(k)}\left(E + \frac{\alpha_1}{n}, E + \frac{\alpha_2}{n}, \ldots, E + \frac{\alpha_k}{n}\right)$$

exists in some sense? If it does, what is the limit?

- *Below the microscopic scale: Wegner estimates.* Let $E \in \text{supp}\,\rho$, f a continuous bounded function and $I = \left[E - \frac{1}{n^\alpha}, E + \frac{1}{n^\alpha}\right]$ for some $\alpha > 1$, is it true that

$$\frac{1}{n\,|I|}\left|\mathbb{E}\sum_{\lambda_i\in I} f\left(\lambda_i\right) - n\int_I f(x)\rho(x)dx\right| \to 0?$$

- *Properties of eigenvectors.* Even though it is not as fundamental as the studies of eigenvalues, there are interesting problems and results involving eigenvectors of random matrix models.
- *Universality.* The famous law of large numbers states that, given a sequence of identical independently distributed (i.i.d.) random variables $\xi_1, \ldots, \xi_n, \ldots$ of mean 0 and finite variance, then the average

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\xi_i$$

almost certainly converges to 0. The exact distribution ξ_1 will not affect the limiting object. A slightly more advanced example is the central limit theorem, which states that given a sequence of i.i.d. random variables $\xi_1, \ldots, \xi_n, \ldots$ of mean 0 and finite variance,

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i$$

converges weakly to the Gaussian distribution. Again, the exact distribution of ξ_i does not matter, as long as it is of mean 0 and variance 1 and the limiting object is *universal* for these distributions.

The general belief is that random matrices belonging to the same "symmetry class"—the Hermitian structure of matrices—behaves similarly, at all levels down to the microscopic level. Let $\mathbf{A}$ be a random Hermitian matrices where the upper triangular entries are i.i.d., mean 0 and variance 1. Let $\tilde{\mathbf{A}}$ be another such random Hermitian matrices, but with a different entry distribution. Do they have the same macroscopic/mesocopic/microscopic behavior? Let $\mathbf{\Lambda}$ and $\tilde{\mathbf{\Lambda}}$ be two eigenvalue ensembles that are invariant under unitary transformation and suitably normalized. Do they have the same macroscopic/mesocopic/microscopic behavior?

- *Fluctuations of global eigenvalue statistics.* As a consequence of the Wigner semicircle law, for any f bounded continuous function, and $\lambda_1, \ldots, \lambda_n$ rescaled eigenvalues of a Wigner matrix, we almost certainly have

$$\frac{1}{n} \sum_{i=1}^{n} f\left(\lambda_i\right) \to \int f(x)\rho_{SC}(x)dx$$

Can we also say anything about the fluctuations of global eigenvalue statistics

$$\frac{1}{n} \sum_{i=1}^{n} f\left(\lambda_i\right) - \mathbb{E}\frac{1}{n} \sum_{i=1}^{n} f\left(\lambda_i\right)?$$

The above quantity is known to converge to a Gaussian random variable of finite variance if f is sufficiently smooth and the variance will diverge if f is an indicator function. It is believed that there is some critical regularity for f for which a central-limit-type theorem holds.

3.3 Change Point of Views: From Vectors to Measures

One of the first problems is to find a mathematically efficient way to express the limit of a vector whose size grows to ∞. Recall that there are n eigenvalues to estimate in our problem and n goes to ∞. A fairly natural way to do this is to associate any vector with a probability measure. More explicitly, suppose we have a vector $\left(y_1, \ldots, y_n\right)$ in $\mathbb{R}^n$. We can associate it with the following measure:

$$dG_n(x) = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}(x)$$

G_n is thus a measure with n point masses of equal weight, one at each of the coordinates of the vector.

We will denote by H_n the spectral distribution of the true (sometimes called population) covariance matrix $\mathbf{\Sigma}_n$, the measure associated with the vector of eigenvalues $\lambda_i, i = 1, \ldots, n$ of $\mathbf{\Sigma}_n$. We will refer to H_n as the true spectral distribution. We can write this measure as

$$dH_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i}(x)$$

where δ_{λ_i} is a point mass, of mass 1, at λ_i. We also call δ_{λ_i} a "dirac" at λ_i. The simplest example of a true spectral distribution is found when $\mathbf{\Sigma}_n = \mathbf{I}_n$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix. In this case, for all i, $\lambda_i = 1$, and $dH_n = \delta_1$. So the true spectral distribution is a point mass at 1 when $\mathbf{\Sigma}_n = \mathbf{I}_n$.

Similarly, we will denote by F_n the measure associated with the eigenvalues ℓ_i, $i = 1, \ldots, n$ of the sample covariance matrix $\mathbf{S}_n$. We refer to F_n as the empirical spectral distribution. Equivalently, we define

$$dF_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\ell_i}(x)$$

The change of focus from vector to measure implies a change of focus in the notion of convergence that we will consider adequate. In particular, for consistency issues, the notion of convergence we will use is weak convergence of probability measures.

3.4 The Stieltjes Transform of Measures

Eigenvalues of a matrix can be viewed as continuous functions of the matrix entries. Nevertheless, these functions do not have closed forms when the matrix size exceeds 4. This is the reason why specific tools are needed for their study. There are three important methods employed in this area: (i) the moment method; (ii) the Stieltjes transform; (iii) orthogonal polynomial decomposition of the exact density of the eigenvalues.

A large number of results concerning the asymptotic properties of the eigenvalues of large dimensional random matrices are formulated in terms of the limiting behavior of the Stieltjes transform of their empirical spectral distributions. The Stieltjes transform is a convenient and very powerful tool in the study of the convergence of spectral distribution of matrices (or operators), just as the characteristic function of a probability distribution is a powerful tool for central limit theorems. Most importantly, there is a simple connection between the Stieltjes transform of the spectral distribution of a matrix and its eigenvalues.

We will consider results obtained via the Stieltjes transform method. By definition, the Stieltjes transform of a measure G on $\mathbb{R}$ is defined as

$$m_G(z) = \int \frac{1}{x - z} dG(x), \quad \text{for } z \in \mathbb{C}^+$$

where

$$\mathbb{C}^+ \triangleq \mathbb{C} \bigcap \{z : \text{Im}\{z\} > 0\}$$

is the set of complex numbers with a strictly positive imaginary part. The Stieltjes transform appears to be known under several names in different areas of mathematics.

It is sometimes referred to as Cauchy or Abel–Stieltjes transform. Good references about Stieltjes transforms include Akhiezer (1965) [167, Sections 3.1–2], Lax (2002) [168, Chapter 32], Hiai and Petz (2000) [169, Chapter 3] and Geronimo and Hill (2003) [170]. Qiu *et al.* [39] also surveyed many properties of the Stieltjes transform.

Here we list important properties of Stieltjes transforms of measures on $\mathbb{R}$:

- If G is a probability measure, $m_G(z) \in \mathbb{C}^+$ if $z \in \mathbb{C}^+$ and $\lim_{y\to\infty} -iym_G(iy) = 1$.
- If F and G are two measures, and if $m_F(z) = m_G(z)$, for all $z \in \mathbb{C}^+$, then $G = F$, almost everywhere [1].
- If G_n is a sequence of probability measures and $m_{G_n}(z)$ has a (pointwise) limit $m(z)$ for all $z \in \mathbb{C}^+$, then there exists a probability measure G with Stieltjes transform $m_G(z) = m(z)$ if and only if $\lim_{y\to\infty} -iym(iy) = 1$. If it is the case, G_n converges weakly to G.
- The same is true if the convergence happens only for an infinite sequence $\{z_i\}_{i=1}^{\infty}$ in $\mathbb{C}^+$ with a limit point in $\mathbb{C}^+$.
- If t is a continuity point of the cumulative distribution function of G, then the derivative $dG(t)/dt = \lim_{\varepsilon\to\infty} \frac{1}{\pi} \operatorname{Im}\left(m_G(t + i\varepsilon)\right)$.

For proofs, we refer the reader to [170].

Through an inversion formula, we can recover the initial measure from its Stieltjes transform $m_{\Gamma_n}(z)$.

The Stieltjes transform characterizes the vague convergence of finite measures. It is an important tool for the study of random matrices.

Let $\mathbb{C}^+ = \{z \in \mathbb{C}, \operatorname{Im}(z) > 0\}$.

Proposition 3.4.1 A sequence $(\mu_n)_{n\geqslant 1}$ of probability measures R converges vaguely to a positive measure μ if and only if their Stieltjes transform $m_{\mu_n}(z)$ for $n \geq 1$ converge to $m_\mu(z)$ on $\mathbb{C}^+$.

The link between the Stieltjes transform and random matrix theory is the following: the Stieltjes transform of the spectral distribution $F_{\mathbf{A}_n}$ of an $n \times n$ matrix $\mathbf{A}_n$ is just

$$m_{\mathbf{A}_n}(z) = \int \frac{1}{x - z} dF_{\mathbf{A}}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\lambda_i - z} = \frac{1}{n} \operatorname{Tr}\left[\left(\mathbf{A}_n - z\mathbf{I}_n\right)^{-1}\right]$$

which is the fundamental relation for studying large dimensional random matrices. Points 3 and 4 above can be used to show *convergence of probability measures* if one can *control the corresponding Stieltjes transforms*. In this sense, the Stieltjes transform plays the role of Fourier transform for a continuous-time (or discrete-time) signal. It is more convenient to work on the problems in the Fourier transformed domain. In analogy, we also study problems in the Stieltjes transformed domain.

1 In measure theory [171], one talks about almost everywhere convergence of a sequence of measurable functions defined on a measurable space. That means pointwise convergence almost everywhere. Suppose $\{f_n\}$ is a sequence of functions sharing the same domain and codomain. The sequence $\{f_n\}$ converges pointwise to f, often written as $\lim_{n\to\infty} f_n \to f$ pointwise, if and only if $\lim_{n\to\infty} f_n(x) \to f(x)$ for every x in the domain.

Like the Fourier transformation in probability theory, there is also a one-to-one correspondence between the distributions and their Stieltjes transforms via the inversion formula: for any distribution function G

$$G\left\{\left[a,b\right]\right\} = \frac{1}{\pi}\lim_{\eta\to\infty}\int_a^b \operatorname{Im} m_G\left(\xi + i\eta\right) d\xi$$

3.5 A Fundamental Result: The Marchenko–Pastur Equation

In the study of covariance matrices, a remarkable result exists that describes the limiting behavior of the empirical spectral distribution, $F_\infty = \lim_{n\to\infty} F_n$, in terms of the limiting behavior of the true spectral distribution, $H_\infty = \lim_{n\to\infty} H_n$. The connection between these two measures F_∞ and H_∞ is made through an equation that links the Stieltjes transform of the empirical spectral distribution to an integral against the true spectral distribution. We call this equation the Marchenko–Pastur equation because it first appeared in the landmark paper of Marchenko and Pastur (1967) [172]. The result was independently rediscovered in Wachter (1978) [173] and then refined in Silverstein and Bai (1995) [174] and Silverstein (1995) [175]. In particular, Silverstein (1995) [175] is the only paper where the case of a nondiagonal true covariance matrix is treated.

We work with the $N \times n$ data matrix $\mathbf{X}$. The sample covariance matrix is defined as

$$\mathbf{S}_n = \frac{1}{N}\mathbf{X}^H\mathbf{X}$$

and denotes m_{F_n} the Stieltjes transform of the spectral distribution, F_n, of $\mathbf{S}_n$. If the data vectors are not made zero mean in the definition of $\mathbf{S}_n$, the difference between two definitions is a matrix of rank one.

In the spectral analysis of $\mathbf{S}_n$, it is usual to assume that the data size p tends to infinity proportionally to the sample size n, or

$$n \to \infty, p \to \infty, \text{ but } \frac{p}{n} \to c \in (0,\infty).$$

When we consider sample covariance matrices $\mathbf{S}_n$, the eigenvalues are random variables, and the corresponding empirical spectral distributions $F_{\mathbf{S}_n}(x)$ are random probability measures on $\mathbb{R}^+ : x \in \mathbb{R}, x > 0$, or, equivalently, a sequence of random variables of measures.

Let $v_{F_n}(z)$ be the Stieltjes transform of the spectral distribution of $\frac{1}{N}\mathbf{X}\mathbf{X}^H$. The v_{F_n} can be expressed by

$$v_{F_n}(z) = -\left(1 - \frac{n}{N}\right)\frac{1}{z} + \frac{n}{N} m_{F_n}(z)$$

Currently, the most general version of the result is found in [175] and states the following:

Theorem 3.5.1 Suppose the data matrix $\mathbf{X}$ can be written a $\mathbf{X} = \mathbf{Y}\boldsymbol{\Sigma}_n^{1/2}$, where $\boldsymbol{\Sigma}_n$ is an $n \times n$ positive definite matrix and $\mathbf{Y}$ is an $N \times n$ matrix whose entries are i.i.d (real or complex), with zero mean and variance 1, $\mathbb{E}\left(Y_{ij}\right) = 0, \mathbb{E}\left(\left|Y_{ij}\right|^2\right) = 1$, and the finite fourth moment $\mathbb{E}\left(\left|Y_{ij}\right|^4\right) < \infty$. Call H_n the true (or population) spectral distribution,

i.e the distribution that puts mass $1/n$ at each of the eigenvalues of the true covariance matrix $\mathbf{\Sigma}_n$. Assume that H_n converges weakly to a limit denoted $H_\infty = \lim_{n\to\infty} H_n$. (We write this convergence $H_n \Rightarrow H_\infty$.) Then, when $n, N \to \infty$, and $n/N \to \gamma, \gamma \in (0, \infty)$, almost certainly where $\nu_\infty(z)$ is a deterministic function;

1) $\nu_{F_n}(z) \to \nu_\infty(z)$
2) $\nu_\infty(z)$ satisfies the equation

$$-\frac{1}{\nu_\infty(z)} = z - \gamma \int \frac{\lambda dH_\infty(\lambda)}{1 + \lambda \nu_\infty(z)}, \quad \forall z \in \mathbb{C}^+ \tag{3.3}$$

3) (3.3) has one and only one solution, which is the Stieltjes transform of a measure.

Theorem 3.5.1 says that the spectral distribution of the sample covariance matrix is asymptotically nonrandom. Furthermore, it is fully characterized by the true population spectral distribution, through (3.3).

Example 3.5.2 (Marchenko–Pastur Law) The white Gaussian random vector has a true covariance matrix $\mathbf{\Sigma}_n = \mathbf{I}_n$; all the population eigenvalues $\lambda_i\left(\mathbf{\Sigma}_n\right)$ are equal to 1. Then, $H_n = H_\infty = \delta_1$. A little bit of elementary work leads to the well known fact in random matrix theory that the empirical spectral distribution, F_n, converges (almost surely) to the Marchenko–Pastur law, if $\gamma < 1$, whose density is given by

$$f_\gamma(x) = \frac{1}{2\pi\gamma x}\sqrt{(b-x)(x-a)}, \quad a = \left(1-\sqrt{\gamma}\right)^2, b = \left(1+\sqrt{\gamma}\right)^2 \tag{3.4}$$

We refer the reader to [172, 176], and [177] for more details and explanations concerning the case $\gamma > 1$. One point of statistical interest is that even though the true population eigenvalues are all equal to 1, the empirical ones are now spread on the interval $\left[\left(1-\sqrt{\gamma}\right)^2, \left(1+\sqrt{\gamma}\right)^2\right]$. □

This Marchenko–Pastur law is the analogue of Wigner's semicircle law in this setting of multiplicative rather than additive symmetrization. The assumption of Gaussian entries may be significantly relaxed.

El Karoui (2008) [178] proposed using a fundamental result in random matrix theory, the Marchenko–Pastur equation (3.3), to better estimate the eigenvalues of large dimensional covariance matrices. The Marchenko–Pastur equation holds in a very wide generality and under weak assumptions. The estimator he obtained can be thought of as "shrinking" in a nonlinear fashion the eigenvalues of the sample covariance matrix to estimate the true population eigenvalue.

3.6 Linear Eigenvalue Statistics and Limit Laws

The empirical spectral density (ESD) of an $n \times n$ Hermitian matrix $\mathbf{A}_n$, which is a one-dimensional function

$$F_{\mathbf{A}_n}(x) = \frac{1}{n}\left|\left\{1 \leqslant i \leqslant n : \lambda_i\left\{\mathbf{A}_n\right\} \leqslant x\right\}\right| = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left(\lambda_i\left\{\mathbf{A}_n\right\} \leqslant x\right) \tag{3.5}$$

where $|I|$ denotes the number of the elements in a finite set I, and $\mathbf{1}(B)$ denotes the indicator of an event B. If the eigenvalues λ_i are not all real, we can define a two-dimensional empirical spectral distribution of the matrix $\mathbf{A}$:

$$F_{\mathbf{A}_n}(x,y) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left(\mathrm{Re}\,\lambda_i\left\{\mathbf{A}_n\right\} \leqslant x, \mathrm{Im}\,\lambda_i\left\{\mathbf{A}_n\right\} \leqslant y\right) \tag{3.6}$$

Sometimes it is more convenient to work with measures than with corresponding distribution functions. We define an empirical spectral measure of eigenvalues of the matrix $\mathbf{A}_n$:

$$\mu_{\mathbf{A}_n}(B) = \frac{1}{n}\left|\left\{1 \leqslant i \leqslant n : \lambda_i\left\{\mathbf{A}_n\right\} \leqslant B\right\}\right|, \quad B \in \mathcal{B}(\mathbb{T})$$

where $\mathbb{T} = \mathbb{R}$ or $\mathbb{T} = \mathbb{C}$ and $\mathcal{B}(\mathbb{T})$ is a Borel σ-algebra[2] of $\mathbb{T}$.

A Wigner matrix is a Hermitian (or symmetric in the real case) matrix in which the upper diagonal and diagonal entries are independent random variables. In this context, we consider the Wigner matrix $\mathbf{M}_n = \{\xi_{ij}\}_{1\leqslant i,j\leqslant n}$, which has the upper diagonal entries as independent, identically distributed (i.i.d.) complex (or real) reandom variables with zero mean and variance 1, and the diagonal entries as i.i.d. real random variables with bounded mean and variance.

A cornerstone of random matrix theory is Wigner's semicircle law. For any real number x, we have

$$\lim_{n\to\infty} F_{\mathbf{A}_n}(x) = \lim_{n\to\infty}\frac{1}{n}\left|\left\{1 \leqslant i \leqslant n : \lambda_i\left\{\mathbf{A}_n\right\} \leqslant x\right\}\right| = \int_{-2}^{x} \rho_{sc}(y)dy \tag{3.7}$$

in the sense of probability (and also in the almost sure sense). There are four types of convergence: almost surely, in law, in probability and in the r-th mean. See, for example, [179] for these definitions.

A fundamental result is Wigner's semicircle law, which describes the global limiting behavior of eigenvalues of the Wigner ensemble: for any bounded continuous function φ, one has

$$\frac{1}{n}\sum_{i=1}^{n}\varphi\left(\lambda_i\right) \underset{\mathbb{P}}{\to} \int \varphi(x)\rho_{sc}(x)\,dx \tag{3.8}$$

where $\rho_{sc}(x) = \frac{1}{2\pi^2}\sqrt{4-x^2}1_{\{|x|\leqslant 2\}}$ is the density function of the Wigner semicircle law $F_{sc}(x)$. We say the empirical spectral density $F_n(x)$ converges *weakly* in probability to the semicircle law $F_{sc}(x)$. The result of this type, which is the analog of the Law of Large Numbers of classical probability theory, is normally the first step in studies of the eigenvalue distribution for any ensemble of random matrices. Central limit theorem (CLT) for fluctuations of linear eigenvalue statistics is a natural second step in studies of the eigenvalue distribution of any ensemble of random matrices (see Section 3.7).

For given random variables X and $X_1, X_2, \ldots$ on a probability space, X_n is said to *converge to X in probability* [179], if and only if the following condition

$$\lim_{n\to\infty}\mathbb{P}\left(\left|X_n - X\right| \geqslant \varepsilon\right) = 0 \tag{3.9}$$

holds for each positive ε, written as $X_n \underset{\mathbb{P}}{\to} X$.

2 A σ-algebra that is related to the topology of a set. The Borel σ-algebra is defined to be the σ-algebra generated by the open sets (or, equivalently, by the closed sets).

On the other hand, the sample covariance matrix plays an fundamental role in statistics. Let $\mathbf{x}$ be a random vector $\mathbf{x} = \left(X_1, \ldots, X_p\right) \in \mathbb{C}^n$ and assume for simplicity that $\mathbf{x}$ is centered (zero mean). Then the true covariance matrix is given by

$$\mathbb{E}\left(\mathbf{x}\mathbf{x}^H\right) = \left(\operatorname{cov}\left(X_i, X_j\right)\right)_{1 \leqslant i,j \leqslant n}$$

Consider N independent samples or realizations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \in \mathbb{C}^n$ of the random vector $\mathbf{x}$ and form the $N \times n$ data matrix $\mathbf{X} = \left(\mathbf{x}_1^T, \ldots, \mathbf{x}_N^T\right)^T \in \mathbb{C}^{N \times n}$. Then the sample covariance matrix is an $n \times n$ non-negative definite matrix defined as

$$\mathbf{S} = \frac{1}{N}\mathbf{X}^H\mathbf{X}$$

If $N \to +\infty$ and n fixed, then the sample covariance matrix converges (entrywise) to the true covariance matrix almost surely. We focus on the regime that both n and N tend to infinity at the same time.

Let $\mathbf{X} = \left\{\xi_{ij}\right\}_{1 \leqslant i \leqslant N, 1 \leqslant j \leqslant n}$ be a random $N \times n$ matrix, where $N = N(n)$ is an integer such that $N \leq n$ and $\lim_{n\to\infty} N/n = y$ for some $y \in (0, 1]$. The matrix ensemble is said to obey condition **C1** with constant C_0 if the random variables ξ_{ij} are jointly independent, having a mean of 0 and variance of 1, and obey the moment condition

$$\sup_{i,j} \mathbb{E}\left|\xi_{ij}\right|^{C_0} \leqslant C$$

for some constant C independent of n, N.

The first fundamental result concerning the asymptotic behavior of empirical spectral density for large covariance matrices is the Marchenko–Pastur law [172, 180–182].

Theorem 3.6.1 (Marchenko–Pastur law) Assume a random $N \times n$ matrix $\mathbf{X}$ obeys condition **C1** with $C_0 \geq 4$, and $n \to \infty$, $N \to \infty$ such that $\lim_{n\to\infty} N/n = y$ for some $y \in (0, 1]$, the empirical spectral distribution of the matrix $\mathbf{S} = \frac{1}{n}\mathbf{X}^H\mathbf{X}$ converges in distribution to the Marchenko–Pastur law with a density function

$$f_{MP}(x) = \frac{1}{2\pi x y \sigma^2}\sqrt{(b-x)(x-a)}\,\mathbb{I}(a \leqslant x \leqslant b) \tag{3.10}$$

where

$$a = \sigma^2\left(1 - \sqrt{y}\right)^2, \quad b = \sigma^2\left(1 + \sqrt{y}\right)^2$$

If $c > 1$, the Marchenko—Pastur law has a point mass $1 - c^{-1}$ at the origin.

Here $\mathbb{I}(\cdot)$ is the indicator function and σ^2 (the variance) is the scale parameter. If $\sigma^2 = 1$, the Marchenko–Pastur law is called the standard Marchenko–Pastur law. When $y = 1$ or $N = n$, the density function is supported on the interval $[0, 4]$ and

$$\frac{d\mu}{dx} = f_{MP}(x) = \frac{1}{2\pi}\sqrt{\frac{4-x}{x}}$$

Actually, by a change of variable $x \to x^2$, the distribution μ is the image of the semicircle law.

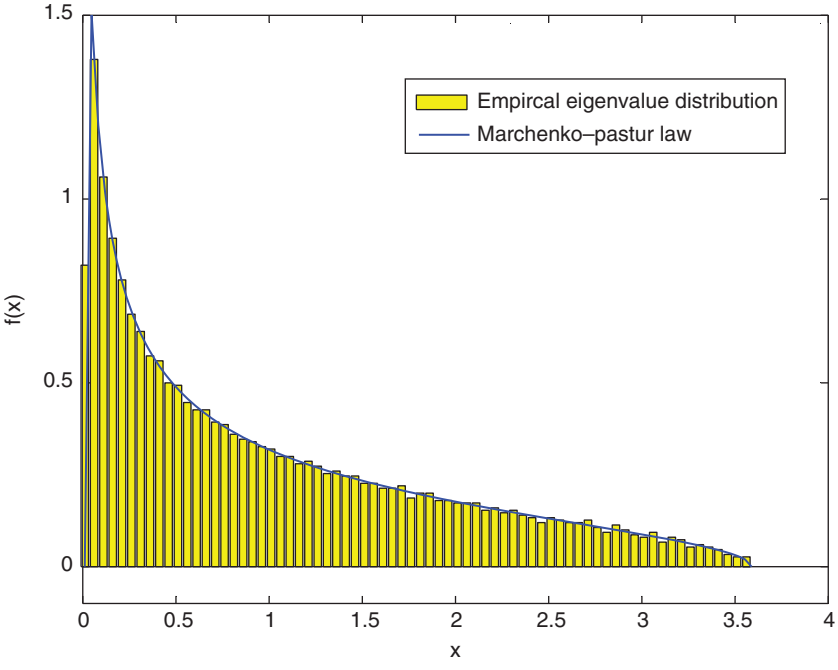


Figure 3.1 Plotted above is the distribution of the eigenvalues of $\frac{1}{n}\mathbf{X}^H\mathbf{X}$ where $\mathbf{X}$ is an $N \times n$ random Gaussian matrix with $n = 3000$ and $y = N/n = 0.8$. The blue curve is the Marchenko–Pastur law with density function $f_{MP}(x)$.

When the aspect ratio $y = N/m = 1$; we get the special case that

$$f(x) = \frac{1}{\pi}\sqrt{4 - x^2}, \quad x \in [0, 2]$$

This is the famous quarter-circle law. The singular values of a normally distributed square matrix lie on a quarter circle. The moments are Catalan numbers.

Figure 3.1 and Figure 3.2 illustrate the Marchenko–Pastur law and compares theoretical predictions with simulations.

The MATLAB codes are shown below.

MATLAB Code for Marchenko–Pastur Law, Revised from [183]

```
%Experiment : Gaussian Random Matrix
%Plot : Histogram of the eigenvalues of XX /m
%Theory : Marcenko-Pastur as n \to infinity
%% Parameters
t =1; %tria s
y = 0.1 ; %aspect ratio
n =3000; %matrix column size
m=round ( n/y ) ;
v = [] ; %eigenvalue samples
dx = 0.05 ; %bin size
```

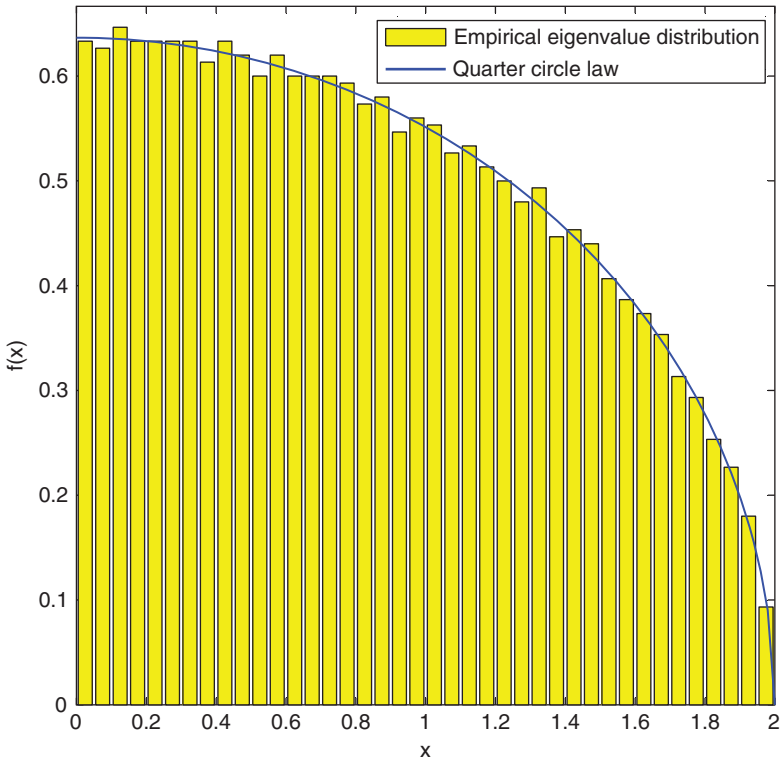


Figure 3.2 Plotted above is the distribution of the eigenvalues of $\frac{1}{n}\mathbf{X}^H\mathbf{X}$ where $\mathbf{X}$ is an $N \times n$ random Gaussian matrix with $n = 3000$ and $y = N/n = 1$. The blue curve is the quarter circle law with density function.

```
%% Experiment
for i =1: t ,
X=randn (m, n ) ; % random m*n matrix
s=X'*X; %symmetric posisitve definite matrix
v =eig (s) ; % eigenvalues
end
v=v/m; % normalized eigenvalues
a=(1-sqrt(y))^2 ; b=(1+ sqrt(y))^2 ;
%% Pl o t
[ count , x ]= hist ( v , a : dx : b ) ;
cla reset
bar ( x , count / ( t*n*dx ) , 'y ' ) ;
hold on ;
%% Theory
x= linspace ( a , b ) ;
plot ( x , sqrt ( ( x-a ) .*( b-x ) ) ./ ( 2 * pi*x*y ) ,
'LineWidth' , 2 )
axis ( [ 0 ceil(b) -0.1 1.5 ] ) ;
xlabel('x')
```

```
ylabel('f(x)')
legend('Empircal Eigenvalue Distribution',
'Marchenko-Patur Law')
```

MATLAB Code for Quarter Circle Law, Revised from [183]

```
%Experime n t : Gaussian Random
%Plot : Histogram singular values
%Theory : Quater Circle Law
%% Parame t e r s
t =1; %trials
r =1; %aspect ratio
n =2000; %matri x column size
m = n ;
v = [ ] ; %eigen value samples
dx = .05 ; %bin size
a = 0 ; b = 2 ;
%% Ex p e r ime n t
for i =1: t ,
v = svd ( randn ( n ) ) ; % singular values
end
v=v / sqrt (m) ; % normalized singular values
close all ;
[ count , x ]= hist ( v , (a-dx/2) : dx : b ) ; cla reset
bar ( x , count / ( t *n*dx ) , ' y ') ; hold on ;
%% Theory
x= linspace (a , b ) ;
plot ( x , sqrt (4 - x.^ 2 ) / pi , 'LineWidth' , 2 )
axis square
axis ( [ 0 2 0 2 / 3 ] ) ;
xlabel('x')
ylabel('f(x)')
legend('Empircal Eigenvalue Distribution',
'Quater Circle Law')
```

Example 3.6.2 (linear vector channel) This example follows [52]. The linear vector memoryless channel is defined as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n} \tag{3.11}$$

where $\mathbf{x} \in \mathbb{C}^K$ is the input vector, $\mathbf{y} \in \mathbb{C}^N$ the output vector and $\mathbf{n} \in \mathbb{C}^N$ models the additive circularly symmetric Gaussian noise. With the i.i.d. Gaussian input, the normalized input-output mutual information of (3.11) conditioned on $\mathbf{H}$ is

$$\begin{aligned}\frac{1}{N} I(\mathbf{x};\mathbf{y}|\mathbf{H}) &= \frac{1}{N}\log\det\left(\mathbf{I} + \mathrm{SNR}\,\mathbf{H}\mathbf{H}^H\right) \\ &= \frac{1}{N}\sum_{i=1}^{N}\log\left(1 + \mathrm{SNR}\,\lambda_i\left(\mathbf{H}\mathbf{H}^H\right)\right) \\ &= \int_0^\infty \log(1+\mathrm{SNR}\,x)\,dF_{\mathbf{H}\mathbf{H}^H}(x)\end{aligned} \tag{3.12}$$

with the transmitted signal-to-noise ratio (SNR)

$$\mathrm{SNR} = \frac{N\mathbb{E}\left[\|\mathbf{x}\|^2\right]}{K\mathbb{E}\left[\|\mathbf{n}\|^2\right]}$$

and with $\lambda_i\left(\mathbf{H}\mathbf{H}^H\right)$ equal to the i-th squared singular value of $\mathbf{H}$. Here $||\cdot||$ denotes the Euclidean norm.

Another fundamental performance measure for (3.11) is the minimum mean-square-error (MMSE) achieved by a linear receiver, which determines the maximum achievable output signal-to-interference-and-noise ratio (SINR). For an i.i.d. input vector, the arithmetic mean over uses (or transmit antennas) of the MMSE is given, as a function of random matrix $\mathbf{H}$, by

$$\begin{aligned}\frac{1}{K}\min_{\mathbf{M}\in\mathbb{C}^{K\times N}}\mathbb{E}\left[\|\mathbf{x}-\mathbf{M}\mathbf{y}\|^2\right] &= \frac{1}{K}\mathrm{Tr}\left\{\left(\mathbf{I}+\mathrm{SNR}\,\mathbf{H}^H\mathbf{H}\right)^{-1}\right\}\\ &= \frac{1}{K}\sum_{i=1}^{K}\frac{1}{1+\mathrm{SNR}\,\lambda_i\left(\mathbf{H}^H\mathbf{H}\right)}\\ &= \int_0^\infty \frac{1}{1+\mathrm{SNR}\,x}dF_{\mathbf{H}^H\mathbf{H}}(x)\\ &= \frac{N}{K}\int_0^\infty \frac{1}{1+\mathrm{SNR}\,x}dF_{\mathbf{H}\mathbf{H}^H}(x)-\frac{N-K}{K}\end{aligned} \tag{3.13}$$

The expectation in the first line is over $\mathbf{x}$ and $\mathbf{n}$. The last line follows the following relation

$$NF_{\mathbf{H}\mathbf{H}^H}(x)-NU(x)=KF_{\mathbf{H}^H\mathbf{H}}(x)-KU(x) \tag{3.14}$$

where $U(x)$ is the unit-step function: $U(x)=0, x<0$; $U(x)=1, x>0$.

Both fundamental performance measures (capacity and MMSE) are coupled through

$$\mathrm{SNR}\,\frac{d}{d\,\mathrm{SNR}}\log_e\det\left(\mathbf{I}+\mathrm{SNR}\,\mathbf{H}\mathbf{H}^H\right)=K-\mathrm{Tr}\left\{\left(\mathbf{I}+\mathrm{SNR}\,\mathbf{H}\mathbf{H}^H\right)^{-1}\right\} \tag{3.15}$$

As seen in (3.12) and (3.13), capacity and MMSE are dictated by the distribution of the empirical (squared) singular value distribution of the random channel matrix $\mathbf{H}$. In the simplest case, the entries of $\mathbf{H}$ are i.i.d. Gaussian. More general cases, such as independent (but not identically distributed) entries or even dependent entries, are of interest in this context. □

Example 3.6.3 (functional averages over Gaussian ensembles) The MIMO channel model is defined similarly to (3.11). The result here can be applied to massive MIMO analysis. See Section 15.3. We repeat the definition to fix a different notation. Denoting the number of transmitting antennas by M and the number of receiving antennas by N, the channel model is

$$\mathbf{y}=\mathbf{H}\mathbf{s}+\mathbf{n} \tag{3.16}$$

where $\mathbf{s}\in\mathbb{C}^M$ is the transmitted vector, $\mathbf{y}\in\mathbb{C}$ is the received vector, $\mathbf{H}\in\mathbb{C}^{N\times M}$ is a complex matrix and $\mathbf{n}\in\mathbb{C}^N$ is the zero mean complex Gaussian vector with independent, equal variance entries. We assume that $\mathbb{E}\left(\mathbf{n}\mathbf{n}^H\right)=\mathbf{I}_N$, where $(\cdot)^H$ denotes the complex

conjugate transpose and $\mathbf{I}_N$ the $N \times N$ identity matrix. It is reasonable to put a power constraint

$$\mathbb{E}\left(\mathbf{n}^H\mathbf{n}\right) = \mathbb{E}\left[\mathrm{Tr}\left(\mathbf{n}\mathbf{n}^H\right)\right] \leqslant P,$$

where P is the total transmitted power. The signal-to-noise ratio, denoted by snr, is defined as the quotient of the signal power and the noise power and in this case is equal to P/N.

Recall that if $\mathbf{A}$ is an $n \times n$ Hermitian matrix then there exists $\mathbf{U}$ unitary and $\mathbf{D} = \mathrm{diag}\left(d_1, \ldots, d_n\right)$ such that $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^H$. Given a continuous function f we define $f(\mathbf{A})$ as

$$f(\mathbf{A}) = \mathbf{U}\,\mathrm{diag}\left(f\left(d_1\right), \ldots, f\left(d_n\right)\right)\mathbf{U}^H$$

Naturally, the simplest example is the one where $\mathbf{H}$ has independent and identically distributed (i.i.d.) Gaussian entries, which constitutes the canonical model for the single user narrow band MIMO channel. It is known that the capacity of this channel is achieved when $\mathbf{s}$ is a vector with complex Gaussian zero mean and covariance $\mathrm{snr}\,\mathbf{I}_M$. See [51, 52] for instance. For the fast fading channel, assuming statistical channel state information at the transmitter, the ergodic capacity is given by

$$\mathbb{E}\left[\log\det\left(\mathbf{I}_N + \mathrm{snr}\,\mathbf{H}\mathbf{H}^H\right)\right] = \mathbb{E}\left[\mathrm{Tr}\log\left(\mathbf{I}_N + \mathrm{snr}\,\mathbf{H}\mathbf{H}^H\right)\right] \tag{3.17}$$

where in the last equality we use the fundamental fact that

$$\log\det(\cdot) = \mathrm{Tr}\log(\cdot) \tag{3.18}$$

We prefer the form of $\mathrm{Tr}\log(\cdot)$ because the trace $\mathrm{Tr}(\cdot)$ is a linear function. The expectation $\mathbb{E}(\cdot)$ is also a linear function. Sometimes it is convenient to exchange the order of $\mathbb{E}$ and $\mathrm{Tr}(\cdot)$ in (3.17):

$$\begin{aligned}\mathbb{E}\left[\log\det\left(\mathbf{I}_N + \mathrm{snr}\,\mathbf{H}\mathbf{H}^H\right)\right] &= \mathbb{E}\left[\mathrm{Tr}\log\left(\mathbf{I}_N + \mathrm{snr}\,\mathbf{H}\mathbf{H}^H\right)\right] \\ &= \mathrm{Tr}\left[\mathbb{E}\log\left(\mathbf{I}_N + \mathrm{snr}\,\mathbf{H}\mathbf{H}^H\right)\right]\end{aligned}$$

The $\mathbb{E}(\mathbf{X})$ can be approximated by the arithmetic average $\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i$ when n "snapshots" of the $p \times p$ random matrix $\mathbf{X}$ are observed. As a result, we reach

$$\begin{aligned}\mathbb{E}\left[\log\det\left(\mathbf{I}_N + \mathrm{snr}\,\mathbf{H}\mathbf{H}^H\right)\right] &= \mathbb{E}\left[\mathrm{Tr}\log\left(\mathbf{I}_N + \mathrm{snr}\,\mathbf{H}\mathbf{H}^H\right)\right] \\ &= \mathrm{Tr}\left[\mathbb{E}\log\left(\mathbf{I}_N + \mathrm{snr}\,\mathbf{H}\mathbf{H}^H\right)\right] \\ &\approx \frac{1}{n}\mathrm{Tr}\left[\sum_{i=1}^{n}\log\left(\mathbf{I}_N + \mathrm{snr}\,\mathbf{H}_i\mathbf{H}_i^H\right)\right]\end{aligned} \tag{3.19}$$

which boils down to the sum of random positive definite Hermitian matrices $\mathbf{H}_i\mathbf{H}_i^H$, $i = 1, \ldots, n$, given the i-th "snapshot" $\mathbf{H}_i$ of the random channel matrix $\mathbf{H}$ that is defined in (15.29). See [40] for a whole chapter on the sum of random matrices. The channel capacity with a finite number of samples can be obtained using (3.19). Note that the Frobenius norm is defined as

$$\|\mathbf{B}\|_F^2 \equiv \mathrm{Tr}\left(\mathbf{B}\mathbf{B}^H\right)$$

In (3.19), if we expand the function $\log\left(\mathbf{I}_N + \mathrm{snr}\,\mathbf{H}_i\mathbf{H}_i^H\right)$ using its Taylor series, we can reduce the problem to the sample moments m_k defined as

$$\hat{m}_k = \frac{1}{M}\mathrm{Tr}\left[\left(\frac{1}{N}\mathbf{H}_i\mathbf{H}_i^H\right)^k\right]$$

for an integer $k \geq 1$. Because the sample moments $\hat{m}_k$ are *consistent estimators* of true moments m_k, it is then natural to use the moment method for the inference of the parameters [53, p. 425]. See Section 8.9.3 for this.

More generally, we can expand a function of a random matrix in the form of $f\left(\mathbf{H}\mathbf{H}^H\right)$ in terms of its Taylor series. We can similarly obtain the true moments m_k. We can use sample moments $\hat{m}_k$ to estimate the true moments.

Another important performance measure is the minimum mean square error (MMSE) achieved by a linear receiver, which determines the maximum achievable output signal-to-interference-and-noise ratio (SINR). For an input vector $\mathbf{x}$ with i.i.d. entries of zero mean and unit variance the MSE at the output of the MMSE receiver is given by

$$\min_{\mathbf{M}\in\mathbb{C}^{M\times N}} \mathbb{E}\left[\|\mathbf{x}-\mathbf{M}\mathbf{y}\|^2\right] = \mathbb{E}\left[\operatorname{Tr}\log\left(\mathbf{I}_M + \operatorname{snr}\mathbf{H}^H\mathbf{H}\right)^{-1}\right] \tag{3.20}$$

where the expectation on the left-hand side is over both the vectors x and the random matrices $\mathbf{H}$, whereas the right-hand side is over $\mathbf{H}$ only. See [52] for details.

Let $\mathbf{H}$ be an $n\times n$ Gaussian random matrix with complex, independent, and identically distributed entries of zero mean and unit variance. Given an $n\times n$ positive definite matrix $\mathbf{A}$, and a continuous function $f:\mathbb{R}^+\to\mathbb{R}$ such that $\int_0^\infty e^{-\alpha t}|f(t)|^2dt < \infty$ for every $\alpha > 0$, Tucci and Vega (2013) [54] find a new formula for the expectation

$$\mathbb{E}\left[\operatorname{Tr}\left(f\left(\mathbf{H}\mathbf{A}\mathbf{H}^H\right)\right)\right]$$

Taking $f(x) = \log(1+x)$ gives another formula for the capacity of the MIMO communication channel, and taking $f(x) = (1+x)^{-1}$ gives the MMSE achieved by a linear receiver.

Let $\mathbb{M}_n$ be the set of all $n\times n$ complex matrices and $\mathbb{U}_n$ be the set of $n\times n$ unitary complex matrices. Let $d\mathbf{H}$ be the Lebesgue measure on $\mathbb{M}_n$ and let

$$d\mu(\mathbf{H}) = \pi^{-n^2}\exp\left(-\operatorname{Tr}\left(\mathbf{H}^H\mathbf{H}\right)\right)d\mathbf{H}$$

be the Gaussian measure on $\mathbb{M}_n$. This is the induced measure by the Gaussian random matrix with complex independent and identically distributed entries with zero mean and unit variance in the set of matrices, when this is represented as an Euclidean space of dimension $2n^2$. Note that this probability measure is left and right invariant under unitary multiplication, or

$$d\mu(\mathbf{H}\mathbf{U}) = d\mu(\mathbf{U}\mathbf{H}) = d\mu(\mathbf{H})$$

for every unitary $\mathbf{U}$.

Let $\mathbf{A}$ be an Hermitian $n\times n$ matrix for $n=2$ with eigenvalues λ_1 and λ_2. If $\lambda_1 \neq \lambda_2$ then

$$\int_{\mathbb{M}_2}\operatorname{Tr}\left[\log\left(\mathbf{I}_2+\mathbf{H}^H\mathbf{A}\mathbf{H}\right)\right]d\mu(\mathbf{H}) = \frac{f_0(\lambda_1)-f_0(\lambda_2)+\lambda_1 f_1(\lambda_2)-\lambda_2 f_1(\lambda_1)}{\lambda_1-\lambda_2}$$

where

$$f_0(\lambda_i) = \int_0^\infty e^{-t}t\lambda_i\log\left(1+t\lambda_i\right)dt, \text{ and } \quad f_1(\lambda_i) = \int_0^\infty e^{-t}\log\left(1+t\lambda_i\right)dt$$

If $\lambda_1 = \lambda_2 = \lambda$ then

$$\int_{\mathbb{M}_2} \mathrm{Tr}\left[\log\left(\mathbf{I}_2 + \lambda \cdot \mathbf{H}^H\mathbf{H}\right)\right] d\mu\left(\mathbf{H}\right) = \int_0^\infty e^{-t}\left[(1+t)\log(1+t\lambda) + \frac{t\lambda(t-1)}{1+t\lambda}\right] dt$$

Analogously, we can compute explicitly the moments for the two-dimensional case. Let $\mathbf{A}$ be an Hermitian 2×2 matrix with eigenvalues λ_1 and λ_2 and let $m \geq 1$. If $\lambda_1 \neq \lambda_2$ then

$$\int_{\mathbb{M}_2} \mathrm{Tr}\left[\left(\mathbf{H}^H\mathbf{A}\mathbf{H}\right)^m\right] d\mu\left(\mathbf{H}\right) = m!\left((m+1)\frac{\lambda_1^{m+1} - \lambda_2^{m+1}}{\lambda_1 - \lambda_2} + \frac{\lambda_1\lambda_2^m - \lambda_2\lambda_1^m}{\lambda_1 - \lambda_2}\right)$$

If $\lambda_1 = \lambda_2 = \lambda$ then

$$\int_{\mathbb{M}_2} \mathrm{Tr}\left[\left(\mathbf{H}^H\mathbf{A}\mathbf{H}\right)^m\right] d\mu\left(\mathbf{H}\right) = m!\left(m^2 + m + 2\right)\lambda^m$$

Let $\mathbf{A}$ be an $n \times n$ positive definite matrix, and let $\left\{\lambda_1, \ldots, \lambda_n\right\}$ be the set of eigenvalues of $\mathbf{A}$. Assume that all the eigenvalues are different. Then

$$\int_{\mathbb{M}_n} \mathrm{Tr}\left[\log\left(\mathbf{I}_n + \mathbf{H}^H\mathbf{A}\mathbf{H}\right)\right] d\mu\left(\mathbf{H}\right) = \frac{1}{\det\left(\mathbf{\Delta}\left(\mathbf{D}\right)\right)} \sum_{k=0}^{n-1} \det\left(\mathbf{T}_k\right) \tag{3.21}$$

where $\mathbf{T}_k$ is the matrix constructed by replacing the $(k+1)$ row of $\mathbf{\Delta}\left(\mathbf{D}\right)$ $\left(\left\{\lambda_i^{n-k-1}\right\}_{i=1}^n\right)$ by

$$\left\{\frac{1}{(n-k-1)!}\int_0^\infty e^{-t}\left(t\lambda_i\right)^{n-k-1}\log\left(1+t\lambda_i\right)\right\}_{i=1}^n$$

Here $\mathbf{\Delta}\left(\mathbf{D}\right)$ is the Vandermonde matrix associated with the sequence $\left\{\lambda_1, \ldots, \lambda_n\right\}$

Let $\mathbf{A}$ be an $n \times n$ positive definite matrix, and let $\left\{\lambda_1, \ldots, \lambda_n\right\}$ be the set of eigenvalues of $\mathbf{A}$. Assume that all the eigenvalues are different. Then

$$\int_{\mathbb{M}_n} \mathrm{Tr}\left[\left(\mathbf{I}_n + \mathbf{H}^H\mathbf{A}\mathbf{H}\right)^{-1}\right] d\mu\left(\mathbf{H}\right) = \frac{1}{\det\left(\mathbf{\Delta}\left(\mathbf{D}\right)\right)} \sum_{k=0}^{n-1} \det\left(\mathbf{T}_k\right) \tag{3.22}$$

where $\mathbf{T}_k$ is the matrix constructed by replacing the $(k+1)$ row of $\mathbf{\Delta}(\mathbf{D})(\{\lambda_i^{n-k-1}\}_{i=1}^n)$ by

$$\left\{\frac{1}{(n-k-1)!}\int_0^\infty e^{-t}\left(t\lambda_i\right)^{n-k-1}\left(1+t\lambda_i\right)^{-1}dt\right\}_{i=1}^n$$

As a consequence of (3.21) and (3.22), we have a new formula for the capacity of the MIMO communication channel and for the MMSE described previously in this example.

For every real value $\alpha > 0$ let us define the following class of functions:

$$L_\alpha^2 := \left\{f : \mathbb{R}^+ \to \mathbb{R} : \text{ measurable such that } \int_0^\infty e^{-\alpha t}|f(t)|^2 dt < \infty\right\} \tag{3.23}$$

This is a Hilbert space with respect to the inner product

$$\langle f, g\rangle_\alpha = \int_0^\infty e^{-\alpha t} f(t)g(t)dt$$

Moreover, polynomials are dense with respect to this norm (see [184, Chapter 10]). Let $\mathcal{F}_\alpha$ be the set of continuous functions in L^2_α and let $\mathcal{F}$ be the intersection of all the $\mathcal{F}_\alpha$

$$\mathcal{F} = \bigcap_\alpha \mathcal{F}_\alpha$$

Note that the family $\mathcal{F}$ is a very rich family of functions. For instance, all functions that do not grow faster than polynomials belong to this family. In particular, $f(t) = \log(1+t) \in \mathcal{F}$.

Let $\mathbf{A}$ be an $n \times n$ positive definite matrix, and let $\{\lambda_1, \ldots, \lambda_n\}$ be the set of eigenvalues of $\mathbf{A}$. Assume that all the eigenvalues are different. Then for every $f \in \mathcal{F}$ we have

$$\int_{\mathbb{M}_n} \mathrm{Tr}\left[f\left(\mathbf{H}^H \mathbf{A} \mathbf{H}\right)\right] d\mu(\mathbf{H}) = \frac{1}{\det(\mathbf{\Delta}(\mathbf{D}))} \sum_{k=0}^{n-1} \det\left(\mathbf{T}_k\right), \tag{3.24}$$

where $\mathbf{\Delta}(\mathbf{D})$ is the Vandermonde matrix associated with the matrix $\mathbf{D} = \mathrm{diag}\left(\lambda_1, \ldots, \lambda_n\right)$ and $\mathbf{T}_k$ is the matrix constructed by replacing the $(k+1)$ row of $\mathbf{\Delta}(\mathbf{D})$ $\left(\left\{\lambda_i^{n-k-1}\right\}_{i=1}^n\right)$ by

$$\frac{1}{(n-k-1)!} \left\{f_k\left(\lambda_i\right)\right\}_{i=1}^n$$

where

$$f_k(x) := \int_0^\infty e^{-t}(tx)^{n-k-1} f(tx)\, dt$$ □

3.7 Central Limit Theorem for Linear Eigenvalue Statistics

Theorems for Wigner's semicircle law and the Marchenko–Pastur law can be viewed as random matrix analogs of the law of large numbers from classical probability theory. Thus a central limit theorem for fluctuations of linear eigenvalue statistics is a natural second step in studies of the eigenvalue distribution of any ensemble of random matrices. Here we only give the result for a sample covariance matrix. We refer to Section B.5 for its application in hypothesis testing.

For each $n \geq 1$, let $\mathbf{A}_n = \frac{1}{n}\mathbf{X}_n^H \mathbf{X}_n$ be a real sample covariance matrix of size n, where $\mathbf{X}_n = \left\{X_{ij}\right\}_{1 \leqslant i,j \leqslant n}$, and $\left\{X_{ij} : 1 \leqslant i,j \leqslant n\right\}$ is a collection of real independent random variables with zero mean and unit variance. The eigenvalues are ordered such that $\lambda_1\left(\mathbf{A}_n\right) \leqslant \lambda_2\left(\mathbf{A}_n\right) \leqslant \cdots \leqslant \lambda_n\left(\mathbf{A}_n\right)$. The test function f from the space $\mathcal{H}_s$ has the norm

$$\|f\|_s^2 = \int (1 + 2|\omega|)^{2s} |F(\omega)|^2 d\omega$$

for some $s > 3/2$, where $F(\omega)$ is the Fourier transform of f defined by

$$F(\omega) = \frac{1}{\sqrt{2\pi}} \int e^{j\omega t} f(t) dt$$

We note that if f is a real-valued function with $f \in \mathcal{H}_s$ for some $s > 3/2$, the both f and its derivative f' are continuous and bounded almost everywhere [185]. In particular, this implies that f is Lipschitz.

Suppose that $\mathbb{E}\left[X_{ij}^4\right] = m_4$ for all $1 \leq i, j \leq n$ and all $n \geq 1$. Assume there exists $\varepsilon > 0$ such that

$$\sup_{n\geqslant 1} \sup_{1\leqslant i,j\leqslant n} \mathbb{E}\left|X_{ij}\right|^{4+\varepsilon} < \infty$$

Let f be a real-valued function with $\|f\|_s < \infty$ for some $s > 3/2$. Then

$$\sum_{i=1}^{n} f\left(\lambda_i\left(\mathbf{A}_n\right)\right) - \mathbb{E}\sum_{i=1}^{n} f\left(\lambda_i\left(\mathbf{A}_n\right)\right) \to \mathcal{N}\left(0, v^2\left[f\right]\right) \tag{3.25}$$

in distribution as $n \to \infty$, where the variance $v^2\left[f\right]$ is a function of f defined by

$$v^2\left[f\right] = \frac{1}{2\pi^2}\int_0^4 \int_0^4 \left(\frac{f(x)-f(y)}{x-y}\right)^2 \frac{\left(4-(x-2)\left(y-2\right)\right)}{\sqrt{4-(x-2)^2}\sqrt{4-\left(y-2\right)^2}} dxdy$$

$$+\frac{m_4-3}{4\pi^2}\left(\int_0^4 \frac{x-2}{\sqrt{4-(x-2)^2}} dx\right)^2$$

For big data, we are interested in the performance of algorithms at different scales of matrix sizes n. The variance of the linear eigenvalue statistics, (3.25), does not grow to infinity in the limit $n \to \infty$ for sufficiently smooth test functions. This points to very effective cancellations between different terms of sum and a rigidity property [186] for the distribution of the eigenvalues.

See also [187] for a recent result. Consider a $N \times n$ matrix

$$\mathbf{Y}_n = \frac{1}{\sqrt{n}}\mathbf{\Sigma}_n^{1/2}\mathbf{X}_n$$

where $\mathbf{\Sigma}_n$ is a nonnegative definite Hermitian matrix and $\mathbf{X}_n$ is a random matrix with i.i.d. real or complex standardized entries. The fluctuations of the linear statistics of the eigenvalues:

$$\mathrm{Tr} f\left(\mathbf{Y}_n\mathbf{Y}_n^H\right) = \sum_{i=1}^{N} f\left(\lambda_i\right), \qquad \lambda_i \text{ eigenvalues of } \mathbf{Y}_n\mathbf{Y}_n^H$$

are shown to be Gaussian, in the regime where both dimensions of matrix $\mathbf{Y}_n$ go to infinity at the same pace and in the case where f is an analytic function. The main improvement with respect to Bai and Silverstein's CLT [188] lies in the fact that Najim (2013) [187] considers general entries with finite fourth moment, but whose fourth cumulant is non-null, i.e. whose fourth moment may differ from the moment of a (real or complex) Gaussian random variable. As a consequence, extra terms proportional to

$$|\nu|^2 = \left|\mathbb{E}\left(X_{11}^n\right)^2\right|^2 \qquad \kappa = \mathbb{E}\left|X_{11}^n\right|^4 - |\nu|^2 - 2$$

appear in the limiting variance and in the limiting bias, which not only depend on the spectrum of matrix $\mathbf{\Sigma}_n$ but also on its eigenvectors.

3.8 Central Limit Theorem for Random Matrix $\mathbf{S}^{-1}\mathbf{T}$

As a generalization of the univariate Fisher statistic, random Fisher matrices are widely used in multivariate statistical analysis, for example for testing the equality of two multivariate population covariance matrices. See Section 8.9.6 for testing equality of multiple covariance matrices.

The asymptotic distributions of several meaningful test statistics depend on the related Fisher matrices. Such Fisher matrices have the form

$$\mathbf{F} = \mathbf{S}_y \mathbf{M} \mathbf{S}_x^{-1} \mathbf{M}^H$$

where $\mathbf{M}$ is a nonnegative and non random Hermitian matrix, and $\mathbf{S}_x$ and $\mathbf{S}_y$ are $p \times p$ sample covariance matrices from two independent samples where the populations are assumed to be centered and normalized (i.e. mean 0, variance 1 and with independent components).

In the large-dimensional context, Zheng (2012) [189] established a central limit theorem for linear spectral statistics of a standard Fisher matrix where the two population covariance matrices are equal: the matrix $\mathbf{M}$ is the identity matrix and $\mathbf{F} = \mathbf{S}_y \mathbf{S}_x^{-1}$. In order to extend the CLT of Zheng (2012) [189] to general Fisher matrices, we first need to establish limit theorems for the spectral (eigenvalues) distribution of the matrix $\mathbf{M}\mathbf{S}_x^{-1}\mathbf{M}^H$, or the matrix $\mathbf{S}_x^{-1}\mathbf{T}$ where $\mathbf{T} = \mathbf{M}^H\mathbf{M}$ is **nonrandom**. In many large-dimensional statistic problems, the deterministic matrix $\mathbf{T}$ is usually not invertible or has eigenvalues close to zero, and it is impossible to base the analysis on the CLT of Bai and Silverstein (2004) [190].

Here we consider the product $\mathbf{S}_x^{-1}\mathbf{T}$ of a general determinist and **nonrandom** Hermitian matrix $\mathbf{T}$ by the inverse $\mathbf{S}_x^{-1}$ of a standard sample covariance matrix, due to Zheng *et al.* (2013) [191].

Consider the hypothesis test

$$\begin{aligned} \mathcal{H}_0 &: \mathbf{S}_y \mathbf{M} \mathbf{S}_x^{-1} \mathbf{M}^H \qquad \mathbf{M} = \mathbf{I}, \\ \mathcal{H}_1 &: \mathbf{S}_y \mathbf{M} \mathbf{S}_x^{-1} \mathbf{M}^H \qquad \mathbf{M} \text{ is arbitrary} \end{aligned}$$

For a $p \times p$ random matrix $\mathbf{A}_n$ with eigenvalues $\lambda_i, i = 1, \ldots, p$ linear spectral statistics of type

$$\frac{1}{p} \sum_{i=1}^{p} f\left(\lambda_i\right) = \operatorname{Tr} f\left(\mathbf{A}\right)$$

for various test functions f are of central importance in the theory of random matrices.

Let $\left\{\mathbf{x}_t\right\}, t = 1, \ldots, n$ be a sequence of independent p-dimensional observations with independent and standardized components, so for $\mathbf{x}_t = \left(X_{tj}\right)$, $\mathbb{E}X_{tj} = 0$ and $\mathbb{E}\left|X_{tj}\right|^2 = 1$. The corresponding sample covariance matrix is

$$\mathbf{S} = \frac{1}{n} \sum_{t=1}^{n} \mathbf{x}_t \mathbf{x}_t^H \tag{3.26}$$

Consider the product matrix

$$\mathbf{S}^{-1}\mathbf{T} = \left(\frac{1}{n} \sum_{t=1}^{n} \mathbf{x}_t \mathbf{x}_t^H \right)^{-1} \mathbf{T} \tag{3.27}$$

where $\mathbf{T}$ is a $p \times p$ non-negative definite and **nonrandom** Hermitian matrix. Notice that we do not ask that $\mathbf{T}$ be invertible.

Assumption 3.8.1 The $p \times n$ observation matrix $\left(X_{tj}\right), t = 1, \ldots, n, j = 1, \ldots, p$ are made with independent elements satisfying $\mathbb{E}X_{tj} = 0, \mathbb{E}\left|X_{tj}\right|^2 = 1$. Moreover, for any $\eta > 0$ and as $p, n \to \infty$

$$\frac{1}{np}\sum_{t=1}^{n}\sum_{j=1}^{p}\mathbb{E}\left[\left|X_{tj}\right|^2 \mathbb{I}_{|X_{ij}| \geq \eta\sqrt{n}}\right] \to 0 \tag{3.28}$$

where $\mathbb{I}_{cdot}$ is the indicator function.

The elements are either all real or all complex and we set an index = 1 or = 2, respectively. In the later case, $\mathbb{E}X_{tj}^2 = 0$ for all t, j.

Assumption 3.8.2 In addition to Assumption 3.8.1, the entries $\left\{X_{tj}\right\}$ have an uniform fourth moment $\mathbb{E}\left|X_{tj}\right|^4 = 1 + \kappa$. Moreover, for any $\eta > 0$ and as $p, n \to \infty$

$$\frac{1}{np}\sum_{t=1}^{n}\sum_{j=1}^{p}\mathbb{E}\left[\left|X_{tj}\right|^4 \mathbb{I}_{|X_{ij}| \geq \eta\sqrt{n}}\right] \to 0 \tag{3.29}$$

Assumption 3.8.3 In addition to Assumption 3.8.1, the entries $\left\{X_{tj}\right\}$ have an uniform fourth moment (not necessarily the same). Moreover, for any $\eta > 0$ and as $p, n \to \infty$

$$\frac{1}{np}\sum_{t=1}^{n}\sum_{j=1}^{p}\mathbb{E}\left[\left|X_{tj}\right|^4 \mathbb{I}_{|X_{ij}| \geq \eta\sqrt{n}}\right] \to 0 \tag{3.30}$$

Assumption 3.8.4 The empirical spectrum density H_n of $\{\mathbf{T}_n\}$ tends to a limit H, which is a probability measure not degenerated to the Dirac mass at 0.

Assumption 3.8.5 In addition to Assumption 3.8.4, the operator norm of $\mathbf{T}$ is bounded when $n, p \to \infty$.

Assumption 3.8.6 The dimension p and the sample size n both tend to infinity such that $p/n \to c \in (0, 1)$.

Theorem 3.8.7 Under Assumptions 3.8.1, 3.8.4 and 3.8.6, with probability 1, the empirical spectrum density $F_n(x)$ of $\mathbf{S}^{-1}\mathbf{T}$ tends to a nonrandom distribution $F_{c,H}$ whose Stieltjes transform $s(z)$ is the unique solution to the equation

$$zm(z) = -1 + \int \frac{tdH(t)}{-z - cz^2 m(z) + t} \tag{3.31}$$

The distribution $F_{c,H}$ is then the limiting spectrum density of $\mathbf{S}^{-1}\mathbf{T}$.

We consider a linear spectral statistics of $\mathbf{S}^{-1}\mathbf{T}$ of form

$$F_n(f) = \int f(x)dF_n(x) = \frac{1}{p}\sum_{i=1}^{p} f(\lambda_i),$$

where $\{\lambda_i\}, i = 1, \ldots, p$ are are the eigenvalues of the matrix $\mathbf{S}^{-1}\mathbf{T}$ and f a given test function. A special feature here is that fluctuations of $F_n(f)$ will not be considered around the limiting spectrum density limit $F_{c,H}(f)$, but around $F_{c_n,H_n}(f)$ a finite sample proxy of $F_{c,H}$ obtained by substituting the parameters (c_n, H_n) to (c, H) in the limiting spectral density. Therefore, we consider the random variable

$$X_n(f) = p\left[F_n(f) - F_{y_n,H_n}(f)\right] = p\int f(x)d\left[F_n - F_{y_n,H_n}\right](x)$$

The statements of the central limit theorem are too technical, and are beyond the scope of this book. See [191] for details.

3.9 Independence for Random Matrices

Our aim in this section is to understand independent random matrices. We need this for the likelihood ratio test for large random matrices. See Section 8.11.

A random matrix phenomenon is an observable phenomenon that can be represented in a matrix form, which under repeated measurements yields different outcomes that are not deterministically predictable. Instead, the outcomes obey certain conditions of statistical regularity. The set of descriptions of all possible outcomes that may occur on observing a matrix random phenomenon is called the *sample space* $\mathcal{S}$.

A *matrix event* is a subset of the sample space $\mathcal{S}$. A measure of the degree of certainty with which a given matrix event will occur when observing a matrix random phenomenon can be found by defining a probability function on subsets of the sample space, $\mathcal{S}$, which assigns a probability to every matrix event according to the three postulates of Kolmogorov [192].

A matrix $\mathbf{X} \in \mathbb{R}^{p\times n}$ consisting of np elements $X_{11}(\cdot), X_{12}(\cdot), \cdots, X_{pn}(\cdot)$ that are real valued functions defined on the sample space $\mathcal{S}$ is a real random matrix, if the range $\mathbb{R}^{p\times n}$ of

$$\begin{pmatrix} X_{11}(\cdot) & \cdots & X_{1n}(\cdot) \\ \vdots & & \vdots \\ X_{p1}(\cdot) & \cdots & X_{pn}(\cdot) \end{pmatrix}$$

consists of Borel sets of np-dimensional real space and, if for each Borel set B of real np-tuples, arranged in a matrix,

$$\begin{pmatrix} X_{11} & \cdots & X_{1n} \\ \vdots & & \vdots \\ X_{p1} & \cdots & X_{pn} \end{pmatrix}$$

in $\mathbb{R}^{p\times n}$, the set

$$\left\{ s \in \mathcal{S} : \quad \begin{pmatrix} X_{11}(s_{11}) & \cdots & X_{1n}(s_{1n}) \\ \vdots & & \vdots \\ X_{p1}(s_{p1}) & \cdots & X_{pn}(s_{pn}) \end{pmatrix} \in B \right\}$$

is an event in $\mathcal{S}$.

A scalar function $f_{\mathbf{X}}(\mathbf{X})$ is such that

(i) $f_{\mathbf{X}}(\mathbf{X}) \geqslant 0$

(ii) $\int_{\mathbf{X}} f_{\mathbf{X}}(\mathbf{X})\, d\mathbf{X} = 1$

(iii) $\mathbb{P}(\mathbf{X} \in \mathcal{A}) = \int_{\mathcal{A}} f_{\mathbf{X}}(\mathbf{X})\, d\mathbf{X}$ where $\mathcal{A}$ is a subset of the space of realizations of $\mathbf{X}$, defines the probability density function (pdf) of the random matrix $\mathbf{X}$.

A scalar function $f_{\mathbf{X},\mathbf{Y}}(\mathbf{X}, \mathbf{Y})$ is such that

(i) $f_{\mathbf{X},\mathbf{Y}}(\mathbf{X}, \mathbf{Y}) \geqslant 0$

(ii) $\mathbb{P}((\mathbf{X}, \mathbf{Y}) \in \mathcal{A}) = \int_{\mathcal{A}} \int_{\mathcal{A}} f_{\mathbf{X},\mathbf{Y}}(\mathbf{X}, \mathbf{Y})\, d\mathbf{X} d\mathbf{Y}$ where $\mathcal{A}$ is a subset of the space of realizations of $(\mathbf{X}, \mathbf{Y})$ defines the joint (bimatrix variate) probability density function of the random matrix $\mathbf{X}$ and $\mathbf{Y}$.

We denote the matrix with p rows and q columns by $\mathbf{A}(p \times q)$. Let the random matrices $\mathbf{X}(p \times n)$ and $\mathbf{Y}(r \times s)$ have the joint pdf $f_{\mathbf{X},\mathbf{Y}}(\mathbf{X}, \mathbf{Y})$. Then

(i) the marginal pdf of $\mathbf{X}$ is defined by

$$f_{\mathbf{X}}(\mathbf{X}) = \int_{\mathbf{Y}} f_{\mathbf{X},\mathbf{Y}}(\mathbf{X}, \mathbf{Y})\, d\mathbf{Y}$$

(ii) the conditional pdf of $\mathbf{X}$ given $\mathbf{Y}$ is defined by

$$f_{\mathbf{X}|\mathbf{Y}}(\mathbf{X}|\mathbf{Y}) = \frac{f_{\mathbf{X},\mathbf{Y}}(\mathbf{X}, \mathbf{Y})}{f_{\mathbf{Y}}(\mathbf{Y})}, \quad f_{\mathbf{Y}}(\mathbf{Y}) > 0$$

where $f_{\mathbf{Y}}(\mathbf{Y})$ is the marginal pdf of $\mathbf{Y}$.

Likewise, we can define the marginal pdf of $\mathbf{X}$, and the conditional pdf of $\mathbf{Y}$, given $\mathbf{X}$.

Two random matrices $\mathbf{X}(p \times n)$ and $\mathbf{Y}(r \times s)$ are *independently* distributed if and only if

$$f_{\mathbf{X},\mathbf{Y}}(\mathbf{X}, \mathbf{Y}) = f_{\mathbf{X}}(\mathbf{X}) f_{\mathbf{Y}}(\mathbf{Y})$$

where $f_{\mathbf{X}}(\mathbf{X})$ and $f_{\mathbf{Y}}(\mathbf{Y})$ are the marginal densities of $\mathbf{X}$ and $\mathbf{Y}$, respectively.

The moment generating function (mgf) of the random matrix $\mathbf{X}(p \times n)$ is defined as

$$M_{\mathbf{X}}(\mathbf{Z}) = \int_{\mathbf{X}} \exp\left(\mathrm{Tr}\left(\mathbf{Z}\mathbf{X}^T\right)\right) f_{\mathbf{X}}(\mathbf{X})\, d\mathbf{X}$$

where $\mathbf{Z}(p \times n)$ is a mgf if and only if it is positive and continuous in a neighborhood of $\mathbf{Z} = \mathbf{0}$, where $M_{\mathbf{X}}(\mathbf{0}) = 1$. In this case, the pdf is determined uniquely by the mgf.

The characteristic function of a random matrix $\mathbf{X}(p \times n)$ is defined as

$$\Phi(\mathbf{Z}) = M_{\mathbf{X}}\left(j\mathbf{Z}\right),$$

where $j = \sqrt{-1}$. The mgf of a bimatrix variate distribution is defined by

$$\begin{aligned} M_{\mathbf{X}_1,\mathbf{X}_2}\left(\mathbf{Z}_1, \mathbf{Z}_2\right) &= \mathbb{E}\left[\exp\left\{\mathrm{Tr}\left(\mathbf{Z}_1\mathbf{X}_1^T\right) + \mathrm{Tr}\left(\mathbf{Z}_2\mathbf{X}_2^T\right)\right\}\right] \\ &= \int_{\mathbf{X}_1} \int_{\mathbf{X}_2} \exp\left\{\mathrm{Tr}\left(\mathbf{Z}_1\mathbf{X}_1^T\right)\right. \\ &\quad \left. + \mathrm{Tr}\left(\mathbf{Z}_2\mathbf{X}_2^T\right)\right\} f_{\mathbf{X}_1,\mathbf{X}_2}\left(\mathbf{X}_1, \mathbf{X}_2\right) d\mathbf{X}_1 d\mathbf{X}_2. \end{aligned}$$

The function $M_{\mathbf{X}_1,\mathbf{X}_2}\left(\mathbf{Z}_1, \mathbf{Z}_2\right)$ is an mgf if and only if it is positive and continuous a neighborhood of $\mathbf{Z}_1 = \mathbf{0}$, and $\mathbf{Z}_2 = \mathbf{0}$,where $M_{\mathbf{X}_1,\mathbf{X}_2}(\mathbf{0}, \mathbf{0}) = 1$. The mgf of the marginal distributions of $\mathbf{X}_i, i = 1, 2$ are given by

$$M_{\mathbf{X}_1}\left(\mathbf{Z}_1\right) = M_{\mathbf{X}_1,\mathbf{X}_2}\left(\mathbf{Z}_1, \mathbf{0}\right)$$

and

$$M_{\mathbf{X}_2}\left(\mathbf{Z}_2\right) = M_{\mathbf{X}_1,\mathbf{X}_2}\left(\mathbf{0},\mathbf{Z}_2\right)$$

respectively. In this case, the joint pdf $f_{\mathbf{X}_1,\mathbf{X}_2}\left(\mathbf{X}_1,\mathbf{X}_2\right)$ is determined uniquely by the mgf.

Example 3.9.1 (Gaussian matrix ensembles) A random real symmetric $N\times N$ matrix $\mathbf{X}$ is said to belong to the Gaussian orthogonal ensemble (GOE) if the diagonal and upper triangular elments are *independently* chosen with p.d.f.s

$$\frac{1}{\sqrt{2\pi}}e^{-x_{ii}^2/2} \text{ and } \frac{1}{\sqrt{\pi}}e^{-x_{ij}^2}$$

respectively. An equivalent construction of GOE matrices is to let $\mathbf{A}$ be an $N\times N$ random matrix of independent standard Gaussians $\mathcal{N}(0,1)$ and to form $\mathbf{X} = \frac{1}{2}\left(\mathbf{A}+\mathbf{A}^T\right)$.

The joint p.d.f. of all the independent elements of $\mathbf{X}$ is

$$\begin{aligned} p(\mathbf{X}) : &= \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi}}e^{-x_{ii}^2/2}\prod_{1\leqslant i\leqslant j\leqslant N}\frac{1}{\sqrt{\pi}}e^{-x_{ij}^2} = A_N\prod_{i,j=1}^{N}e^{-x_{ij}^2} \\ &= A_N\exp\left(-\sum_{i,j=1}^{N}x_{ii}^2/2\right) = A_N\exp\left(-\tfrac{1}{2}\operatorname{Tr}\mathbf{X}^2\right) \end{aligned}$$

where A_N is the normalization. The invariance

$$p\left(\mathbf{U}^{-1}\mathbf{X}\mathbf{U}\right) = p(\mathbf{X})$$

for any unitary matrix $\mathbf{U}$, i.e., $\mathbf{U}^H\mathbf{U} = \mathbf{I}$.

Now consider another GOE matrix $\mathbf{Y}$. The joint p.d.f. of all the independent elements of $\mathbf{Y}$ is

$$\begin{aligned} p(\mathbf{Y}) : &= \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi}}e^{-y_{ii}^2/2}\prod_{1\leqslant i\leqslant j\leqslant N}\frac{1}{\sqrt{\pi}}e^{-y_{ij}^2} = B_N\prod_{i,j=1}^{N}e^{-y_{ij}^2} \\ &= B_N\exp\left(-\sum_{i,j=1}^{N}y_{ii}^2/2\right) = B_N\exp\left(-\frac{1}{2}\operatorname{Tr}\mathbf{Y}^2\right) \end{aligned}$$

Now we assume that $\mathbf{X}$ and $\mathbf{Y}$ are independent: all the elements X_{ij} of $\mathbf{X}$ are independent from all the elements Y_{ij} of $\mathbf{Y}$. The joint p.d.f. of all the independent elements of $\mathbf{X}$ and $\mathbf{Y}$ is

$$\begin{aligned} p(\mathbf{X})p(\mathbf{Y}) &= \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi}}e^{-x_{ii}^2/2}\prod_{1\leqslant i\leqslant j\leqslant N}\frac{1}{\sqrt{\pi}}e^{-x_{ij}^2}\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi}}e^{-y_{ii}^2/2}\prod_{1\leqslant i\leqslant j\leqslant N}\frac{1}{\sqrt{\pi}}e^{-y_{ij}^2} \\ &= C_N\prod_{i,j=1}^{N}e^{-x_{ij}^2}\prod_{i,j=1}^{N}e^{-y_{ij}^2} = C_N\exp\left(-\sum_{i,j=1}^{N}x_{ii}^2/2\right)\exp\left(-\sum_{i,j=1}^{N}y_{ii}^2/2\right) \\ &= C_N\exp\left(-\frac{1}{2}\operatorname{Tr}\left(\mathbf{X}^2+\mathbf{Y}^2\right)\right) \end{aligned}$$

□

Example 3.9.2 (Wishart random matrices) The matrix

$$\mathbf{G} = \begin{bmatrix} \mathbf{0}_{n\times n} & \mathbf{X} \\ \mathbf{X}^H & \mathbf{0}_{m\times m} \end{bmatrix} \tag{3.32}$$

where $\mathbf{X}$ is an $n \times m$ $(n \geq m)$ matrix, has in general $n - m$ zero eigenvalues and the remaining eigenvalues given by $\pm$ the positive square roots of the eigenvalues of $\mathbf{X}^H\mathbf{X}$.

Let $\mathbf{X}$ denote an $n \times m$ $(n \geq m)$ random matrix, and suppose the elements of $\mathbf{X}$ are determined by a parameter $\beta = 1, 2$ or 4. These elements are real, complex, and real quaternion independent random variables with Gaussian densities

$$\frac{1}{\sqrt{2\pi}} e^{-x_{ij}^2}, \frac{1}{\pi} e^{-|z|_{ij}^2}, \frac{2}{\pi} e^{-2|z|_{ij}^2} \text{ and } \frac{2}{\pi} e^{-2|w|_{ij}^2}$$

in the three cases $\beta = 1, 2$ or 4. A real quaternion is specified by two complex numbers z and w. Use $\mathbf{X}$ to form $\mathbf{G}$, according to (3.32).

We define Wishart ensembles as consisting of $\mathbf{X}^H\mathbf{X}$, referred to as (uncorrelated) Wishart matrices.

The joint probability density function of the elements of the $n \times m$ complex matrix $\mathbf{X}$ is

$$p(\mathbf{X}) = \frac{1}{\pi^{nm}} \prod_{i=1}^{n} \prod_{j=1}^{m} e^{-|z|_{ij}^2} = \frac{1}{\pi^{nm}} \exp\left(-\operatorname{Tr} \mathbf{X}^H\mathbf{X}\right) \tag{3.33}$$

Similarly , the joint probability density function of the elements of the $n \times m$ complex matrix $\mathbf{Y}$ is

$$p(\mathbf{Y}) = \frac{1}{\pi^{nm}} \prod_{i=1}^{n} \prod_{j=1}^{m} e^{-|w|_{ij}^2} = \frac{1}{\pi^{nm}} \exp\left(-\operatorname{Tr} \mathbf{Y}^H\mathbf{Y}\right) \tag{3.34}$$

When $\mathbf{X}$ and $\mathbf{Y}$ are independent: all the elements of $\mathbf{X}$ are independent of all the elements of $\mathbf{Y}$, the joint p.d.f. of all the elements of X and $\mathbf{Y}$ is given by

$$\begin{aligned} p(\mathbf{X}) p(\mathbf{Y}) &= \frac{1}{\pi^{nm}} \frac{1}{\pi^{nm}} \left(\prod_{i=1}^{n} \prod_{j=1}^{m} e^{-|z|_{ij}^2} \right) \left(\prod_{i=1}^{n} \prod_{j=1}^{m} e^{-|w|_{ij}^2} \right) \\ &= \frac{1}{\pi 2^{nm}} \exp\left(-\operatorname{Tr} \mathbf{X}^H\mathbf{X}\right) \exp\left(-\operatorname{Tr} \mathbf{Y}^H\mathbf{Y}\right) \\ &= \frac{1}{\pi^{2nm}} \exp\left(-\operatorname{Tr}\left(\mathbf{X}^H\mathbf{X} + \mathbf{Y}^H\mathbf{Y}\right)\right) \end{aligned} \tag{3.35}$$

Let us generalize the above two independent random matrices to N $(N \geq 2)$ independent random matrices $\mathbf{X}_i, i = 1, 2, \ldots, N$. We have

$$p\left(\mathbf{X}_1\right) p\left(\mathbf{X}_2\right) \cdots p\left(\mathbf{X}_N\right) = \frac{1}{\pi^{Nnm}} \exp\left(- \operatorname{Tr}\left(\sum_{i=1}^{N} \mathbf{X}_i^H \mathbf{X}_i \right) \right) \tag{3.36}$$

With $\mathbf{X}$ an $n \times m$ complex Gaussian matrix given by (3.34) and $\mathbf{A} = \mathbf{X}^H\mathbf{X}$, we have

$$p(\mathbf{A}) = \int \delta\left(\mathbf{A} - \mathbf{X}^H\mathbf{X}\right) p(\mathbf{X})\, d\mathbf{X} \tag{3.37}$$

Here $\delta\left(\mathbf{A} - \mathbf{X}^H\mathbf{X}\right)$ is equal to the product of one-dimensional delta functions over the independent real and imaginary parts of $\mathbf{A}$. Writing each of these as a Fourier integral shows

$$\delta\left(\mathbf{A} - \mathbf{X}^H\mathbf{X}\right) = \frac{1}{(2\pi)^{m^2}} \int e^{i \operatorname{Tr}\left(\mathbf{H}\left(\mathbf{A} - \mathbf{X}^H\mathbf{X}\right)\right)} d\mathbf{H} \tag{3.38}$$

where $\mathbf{H}$ is an $m \times m$ Hermitian matrix. Substituting this into (3.37) and noting that

$$\int \exp\left(-\operatorname{Tr}\mathbf{X}^H\mathbf{X}\right) e^{i\operatorname{Tr}\left(\mathbf{H}\left(\mathbf{A}-\mathbf{X}^H\mathbf{X}\right)\right)} d\mathbf{X} = \pi^{nm}\left(\det\left(\mathbf{I}+i\mathbf{H}\right)\right)^{-n}$$

and then separating the integration into a product of one-dimensional integrations, gives

$$p\left(\mathbf{A}\right) = \frac{\pi^{2nm}}{\left(2\pi\right)^{m^2}} \int \frac{e^{i\operatorname{Tr}\left(\mathbf{HA}\right)}}{\left(\det\left(\mathbf{I}+i\mathbf{H}\right)\right)^{n}} d\mathbf{H}$$

After some algebra, we obtain the p.d.f. of $\mathbf{A} = \mathbf{X}^H\mathbf{X}$ as

$$p\left(\mathbf{A}\right) = \frac{1}{C_{\beta,N}} \exp\left(-\frac{\beta}{2}\operatorname{Tr}\mathbf{A}\right) \left(\det\mathbf{A}\right)^{\beta/2(n-m+1-2/\beta)} \tag{3.39}$$

where $C_{\beta,N}$ is a normalization constant.

For a rectangular $n \times m$ matrix $\mathbf{X}$ with $n < m$, sometimes we want to deal with the square equivalent $\mathbf{Y}$, where the $m \times m$ $\mathbf{Y}$ is obtained from $\mathbf{X}$ by the addition of $m - n$ rows of zeros. It can be shown that

$$\mathbf{X}^H\mathbf{X} = \mathbf{Y}^H\mathbf{Y}$$

where the $m \times m$ matrix $\mathbf{X}^H\mathbf{X}$ has $m - n$ zero eigenvalues. The nonzero eigenvalues of $\mathbf{X}^H\mathbf{X}$ and $\mathbf{X}\mathbf{X}^H$ are equal.

Consider the correlated Wishart matrices. The $n \times m$ data matrix $\mathbf{X}$ has rows in $\mathbf{X}^T\mathbf{X}$ has each row drawn from an m-dimensional Gaussian with mean zero and variance $\mathbf{\Sigma}$. Equivalently, the distribution of $\mathbf{X}$ is proportional to

$$p\left(\mathbf{X}\right) \propto \exp\left(-\frac{1}{2}\operatorname{Tr}\left(\mathbf{X}^T\mathbf{X}\mathbf{\Sigma}^{-1}\right)\right) \tag{3.40}$$

For the complex case, we have

$$p\left(\mathbf{X}\right) \propto \exp\left(-\frac{1}{2}\operatorname{Tr}\left(\mathbf{X}^H\mathbf{X}\mathbf{\Sigma}^{-1}\right)\right)$$ □

Example 3.9.3 (Probability density functions of random matrices) Assume that the matrix $\mathbf{X} \in \mathbb{C}^{n\times n}$ has a probability density function

$$p_n\left(\mathbf{X}\right) \triangleq H\left(\lambda_1, \ldots, \lambda_n\right)$$

It is known that the joint probability density function of the eigenvalues (which are not necessarily independent, in general) will be of the form

$$p_n\left(\lambda_1, \ldots, \lambda_n\right) = cJ\left(\lambda_1, \ldots, \lambda_n\right) H\left(\lambda_1, \ldots, \lambda_n\right)$$

where $J(\cdot)$ arises from the integral of the Jacobian of the transform from the matrix space to its eigenvalues-eigenvector space, and c is a constant for normalization to make sure that the integral of $p_n\left(\mathbf{X}\right)$ is one. Generally, it is assumed that $H(\cdot)$ has the form

$$H\left(\lambda_1, \ldots, \lambda_n\right) = \prod_{i=1}^{n} g\left(\lambda_i\right) \tag{3.41}$$

and $J(\cdot)$ has the form

$$J\left(\lambda_1, \ldots, \lambda_n\right) = \prod_{i<j}\left(\lambda_i - \lambda_j\right)^{\beta} \prod_{i=1}^{n} h_n\left(\lambda_i\right) \tag{3.42}$$

For example, $\beta = 1$ and $h_n = 1$ for a real Gaussian matrix, $\beta = 2, h_n = 1$ for a complex Gaussian matrix, $\beta = 4, h_n = 1$ for a quaternion Gaussian matrix, and $\beta = 1, h_n = x^{n-p}$ for a real Wishart matrix with $n \geq p$.

Examples are summarized here:

1) Real Gaussian matrix (symmetric; i.e., $\mathbf{X}^T = \mathbf{X}$)

$$p_n(\mathbf{X}) = c \exp\left(-\frac{1}{4\sigma^2} \operatorname{Tr}\left(\mathbf{X}^2\right)\right)$$

The diagonal entries of $\mathbf{X}$ are i.i.d. real $\mathcal{N}\left(0, 2\sigma^2\right)$ and entries above diagonal are i.i.d. real $\mathcal{N}\left(0, \sigma^2\right)$.

2) Complex Gaussian matrix (Hermitian; i.e., $\mathbf{X}^H = \mathbf{X}$)

$$p_n(\mathbf{X}) = c \exp\left(-\frac{1}{2\sigma^2} \operatorname{Tr}\left(\mathbf{X}^2\right)\right)$$

The diagonal entries of $\mathbf{X}$ are i.i.d. real $\mathcal{N}\left(0, \sigma^2\right)$ and entries above diagonal are i.i.d. complex $\mathcal{N}\left(0, \sigma^2\right)$ (their real and imaginary parts are i.i.d. $\mathcal{N}\left(0, \sigma^2/2\right)$).

3) Real Wishart matrix of order $p \times n$

$$p_n(\mathbf{X}) = c \exp\left(-\frac{1}{2\sigma^2} \operatorname{Tr}\left(\mathbf{X}^T\mathbf{X}\right)\right)$$

The entries of $\mathbf{X}$ are i.i.d. real $\mathcal{N}\left(0, \sigma^2\right)$

4) Complex Wishart matrix of order $p \times n$

$$p_n(\mathbf{X}) = c \exp\left(-\frac{1}{\sigma^2} \operatorname{Tr}\left(\mathbf{X}^H\mathbf{X}\right)\right) \tag{3.43}$$

The entries of $\mathbf{X}$ are i.i.d. complex $\mathcal{N}\left(0, \sigma^2\right)$

For generalized densities, we have:

1) The real Gaussian matrix (symmetric; i.e., $\mathbf{X}^T = \mathbf{X}$)

$$p_n(\mathbf{X}) = c \exp\left(-\operatorname{Tr} V(\mathbf{X})\right)$$

The diagonal entries of $\mathbf{X}$ are i.i.d. real $\mathcal{N}\left(0, 2\sigma^2\right)$ and entries above diagonal are i.i.d. real $\mathcal{N}\left(0, \sigma^2\right)$.

2) The complex Gaussian matrix (Hermitian; i.e., $\mathbf{X}^H = \mathbf{X}$)

$$p_n(\mathbf{X}) = c \exp\left(-\operatorname{Tr} V(\mathbf{X})\right).$$

The diagonal entries of $\mathbf{X}$ are i.i.d. real $\mathcal{N}\left(0, \sigma^2\right)$ and entries above diagonal are i.i.d. complex $\mathcal{N}\left(0, \sigma^2\right)$ (the real and imaginary parts of which are i.i.d. $\mathcal{N}\left(0, \sigma^2/2\right)$.

3) The real Wishart matrix of order $p \times n$

$$p_n(\mathbf{X}) = c \exp\left(-\operatorname{Tr} V\left(\mathbf{X}^T\mathbf{X}\right)\right)$$

The entries of $\mathbf{X}$ are i.i.d. real $\mathcal{N}\left(0, \sigma^2\right)$.

4) The complex Wishart matrix of order $p \times n$

$$p_n(\mathbf{X}) = c \exp\left(-\operatorname{Tr} V\left(\mathbf{X}^H\mathbf{X}\right)\right) \tag{3.44}$$

The entries of $\mathbf{X}$ are i.i.d. complex $\mathcal{N}\left(0, \sigma^2\right)$.

In Case 1 and Case 2, $V(x)$ is assumed to be a polynomial of *even* degrees with a positive leading coefficients. For example, we have the $2m$-order polynomial

$$V(x) = \gamma_{2m}x^{2m} + \cdots + \gamma_0, \quad \gamma_{2m} > 0$$

In Case 3 and Case 4, $V(x)$ is assumed to be a polynomial with positive leading coefficients. For example, we have $V(x) = ax^2 + bx + c, \quad a > 0$, where $a > 0, b$ and c are the coefficients for the second-order polynomial. □

Example 3.9.4 (independence for random matrices) The aim of this example is to understand, if two random matrices X and Y are jointly independent, what happens to their probability distribution functions.

Matrix-valued random variables or random matrices take values in a matrix space $\mathbb{M}_{n\times p}(\mathbb{R})$ or $\mathbb{M}_{n\times p}(\mathbb{C})$ of $n \times p$ real or complex-valued matrices, with Borel σ-algebra, where $n, p \geq 1$ are integers. One can view a matrix-valued random variable $X = \left(X_{ij}\right)_{1\leqslant i\leqslant n;1\leqslant j\leqslant p}$ as the *joint random variable* of its scalar components X_{ij}. One can apply all the usual matrix operators (e.g., sum, product, determinant, trace, inverse) on random matrices to obtain a random variable with the appropriate range.

Given a random variable X, taking values in some range R, we define the *distribution* μ_X of X to be the probability measure of the measurable space R defined by the formula

$$\mu_X(S) = \mathbb{P}(X \in S)$$

The distribution of a discrete random variable can be expressed as the sum of Dirac masses

$$\mu_X = \sum_{x\in R} p_x \delta_x$$

Given any $n \times n$ Hermitian matrix M_n, the empirical spectral distribution (or ESD)

$$\mu_{\frac{1}{\sqrt{n}}M_n} := \frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i\left(M_n/\sqrt{n}\right)}$$

of M_n, where

$$\lambda_1\left(M_n\right) \leqslant \cdots \leqslant \lambda_n\left(M_n\right)$$

are the (necessarily real) eigenvalues of M_n, counting multiplicity. The ESD is a probability measure, which can be viewed as a distribution of the normalized eigenvalues of M_n.

When M_n is a random variable ensemble, then the ESD $\mu_{\frac{1}{\sqrt{n}}M_n}$ is now a *random* measure: a random variable taking values in the space $\Pr(\mathbb{R})$ of probability measures on the real line. Thus, the distribution $\mu_{\frac{1}{\sqrt{n}}M_n}$ is a probability measure of probability measures!

A family $\left(X_\alpha\right)_{\alpha\in A}$ of random variables is said to be *jointly independent* if the distribution of $\left(X_\alpha\right)_{\alpha\in A}$ is the product measure of the distribution of the individual X_α.

We say that X is independent of Y if (X, Y) are jointly independent.

A family of events $\left(E_\alpha\right)_{\alpha\in A}$ is said to be jointly independent if their indicators $\left(\mathbb{I}\left(E_\alpha\right)\right)_{\alpha\in A}$ are jointly independent.

A finite family $\left(X_1, \ldots, X_k\right)$ of random variables $X_i, 1 \leq i \leq k$ taking values in measurable spaces R_i are jointly independent **if and only if**

$$\mathbb{P}\left(X_i \in E_i \text{ for all } 1 \leqslant i \leqslant k\right) = \prod_{i=1}^{k} \mathbb{P}\left(X_i \in E_i\right) \tag{3.45}$$

for all measurable $E_i \subset R_i$. In particular, $X_i, i = 1, \ldots, k$ may be matrix-valued random variables.

If $E_1, \ldots, E_k$ are jointly independent events, we have

$$\mathbb{P}\left(\bigcap_{i=1}^{k} E_i\right) = \prod_{i=1}^{k} \mathbb{P}\left(E_i\right) \tag{3.46}$$

and

$$\mathbb{P}\left(\bigcup_{i=1}^{k} E_i\right) = 1 - \prod_{i=1}^{k} \mathbb{P}\left(E_i\right)$$

Let $\left(X_\alpha\right)_{\alpha \in A}$ be a family of random variables (not necessarily independent or finite), and let μ be a probability measure on a measurable space R, and let B be an arbitrary set. Then, after extending the sample space if necessary, one can find an i.i.d. family $\left(Y_\beta\right)_{\beta \in B}$ with distribution μ that is independent of $\left(X_\alpha\right)_{\alpha \in A}$.

For instance, one can create arbitrarily large i.i.d. families of Bernoulli random variables, Gaussian random variables, and so forth, regardless of what other random variables are in play. □

3.10 Matrix-Valued Gaussian Distribution

Theorem 3.10.1

- For $\mathbf{A}\,(m \times m)$ and $\mathbf{B}\,(n \times n)$ $\det(\mathbf{A} \otimes \mathbf{B}) = (\det \mathbf{A})^n (\det \mathbf{B})^m$.
- For $\mathbf{A}\,(m \times m)$ and $\mathbf{B}\,(m \times m)$, $\operatorname{Tr}(\mathbf{A} \otimes \mathbf{B}) = (\operatorname{Tr} \mathbf{A})(\operatorname{Tr} \mathbf{B})$.
- For $\mathbf{A}\,(m \times n), \mathbf{B}\,(p \times q), \mathbf{C}\,(n \times r)$, and $\mathbf{D}\,(q \times s)$, $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD})$.
- For nonsingular matrices $\mathbf{A}$ and $\mathbf{B}$, $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$.

Theorem 3.10.2 For $\mathbf{A}\,(p \times m), \mathbf{B}\,(n \times q), \mathbf{C}\,(q \times m), \mathbf{D}\,(q \times n), \mathbf{E}\,(m \times m)$, and $\mathbf{X}\,(m \times n)$, we have

- $\operatorname{vec}(\mathbf{AXB}) = \left(\mathbf{B}^T \otimes \mathbf{A}\right) \operatorname{vec}(\mathbf{X})$
- $\operatorname{Tr}(\mathbf{CXB}) = \left(\operatorname{vec}\left(\mathbf{C}^T\right)\right)^T \left(\mathbf{I}_q \otimes \mathbf{X}\right) \operatorname{vec}(\mathbf{B})$
- $\operatorname{Tr}\left(\mathbf{DX}^T\mathbf{EXB}\right) = (\operatorname{vec}(\mathbf{X}))^T \left(\mathbf{D}^T\mathbf{B}^T \otimes \mathbf{E}\right) \operatorname{vec}(\mathbf{X})$
  $= (\operatorname{vec}(\mathbf{X}))^T \left(\mathbf{BD} \otimes \mathbf{E}^T\right) \operatorname{vec}(\mathbf{X})$

The random variable $\mathbf{X}$, with the pdf

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\},\ x \in \mathbb{R} \tag{3.47}$$

where $\mu \in \mathbb{R}$ is said to have a normal (or Gaussian) distribution with mean μ and variance σ^2. The multivariate generation of (3.47) for $\mathbf{x} = \left(X_1, \ldots, X_p\right)^T$ is

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}} \frac{1}{\sqrt{\det \boldsymbol{\Sigma}}} \exp\left\{-\frac{1}{2} \operatorname{Tr} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^T\right\} \tag{3.48}$$

where $\mathbf{x} \in \mathbb{R}^p, \mathbf{m} \in \mathbb{R}^p, \boldsymbol{\Sigma} > 0$, and the random vector and the random vector $\mathbf{x}$ is said to have a multivariate normal (or Gaussian) distribution, denoted by $\mathbf{x} \sim \mathcal{N}_p(\mathbf{m}, \boldsymbol{\Sigma})$, with mean vector $\mathbf{m}$ and covariance matrix $\boldsymbol{\Sigma}$.

The random matrix $\mathbf{X}(p \times n)$ is said to have a matrix valued normal (or Gaussian) distribution with mean matrix $\mathbf{M}(p \times n)$ and covariance matrix $\boldsymbol{\Sigma} \otimes \boldsymbol{\Psi}$ where $\boldsymbol{\Sigma}(p \times p) > 0$ and $\boldsymbol{\Psi}(n \times n) > 0$, if $\operatorname{vec}\left(\mathbf{X}^T\right) \sim \mathcal{N}_{pn}\left(\operatorname{vec}\left(\mathbf{M}^T\right), \boldsymbol{\Sigma} \otimes \boldsymbol{\Psi}\right)$. For a matrix $\mathbf{Y}(m \times n)$, $\operatorname{vec}(\mathbf{Y})$ is an $mn \times 1$ vector defined as

$$\operatorname{vec}(\mathbf{Y}) = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_m \end{pmatrix}$$

where $\mathbf{y}_i, i = 1, \ldots, n$ is the i-th column of $\mathbf{Y}$. The notation $\mathbf{A} \otimes \mathbf{B}$ denotes the Kronecker product (direct product) of the matrices $\mathbf{A}$ and $\mathbf{B}$.

We use the notation $\mathbf{X} \sim \mathcal{N}_{p,n}(\mathbf{M}, \boldsymbol{\Sigma} \otimes \boldsymbol{\Psi})$.

If $\mathbf{X} \sim \mathcal{N}_{p,n}(\mathbf{M}, \boldsymbol{\Sigma} \otimes \boldsymbol{\Psi})$, then the pdf of $\mathbf{X}$ is given by

$$p(\mathbf{X}) = \frac{1}{(2\pi)^{np/2}} \frac{1}{(\det \boldsymbol{\Sigma})^{n/2}} \exp\left\{-\frac{1}{2} \operatorname{Tr}\left[\boldsymbol{\Sigma}^{-1} (\mathbf{X} - \mathbf{M}) \boldsymbol{\Psi}^{-1} (\mathbf{X} - \mathbf{M})^T\right]\right\} \tag{3.49}$$

where $\mathbf{X} \in \mathbb{R}^{p \times n}, \mathbf{M} \in \mathbb{R}^{p \times n}$.

We now derive the density of the random matrix $\mathbf{X}$. Let $\mathbf{x} = \operatorname{vec}\left(\mathbf{X}^T\right)$ and $\mathbf{m} = \operatorname{vec}\left(\mathbf{M}^T\right)$. Then, using the definition of (3.49), $\mathbf{x} \sim \mathcal{N}_{p,n}(\mathbf{m}, \boldsymbol{\Sigma} \otimes \boldsymbol{\Psi})$, and its pdf is

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{np/2}} \frac{1}{(\det \boldsymbol{\Sigma} \otimes \boldsymbol{\Psi})^{1/2}} \exp\left\{-\frac{1}{2} \operatorname{Tr}\left[(\boldsymbol{\Sigma} \otimes \boldsymbol{\Psi})^{-1} (\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^T\right]\right\}$$

Using Theorems 3.10.1 and 3.10.2, we get

$$(\det \boldsymbol{\Sigma} \otimes \boldsymbol{\Psi})^{-1/2} = (\det \boldsymbol{\Sigma})^{-n/2} (\det \boldsymbol{\Psi})^{-p/2} \tag{3.50}$$

$$\begin{aligned} \operatorname{Tr}\left[(\boldsymbol{\Sigma} \otimes \boldsymbol{\Psi})^{-1} (\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^T\right] &= \operatorname{Tr}\left[\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Psi}^{-1} (\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^T\right] \\ &= \operatorname{Tr}\left[\boldsymbol{\Sigma}^{-1} (\mathbf{X} - \mathbf{M}) \boldsymbol{\Psi}^{-1} (\mathbf{X} - \mathbf{M})^T\right] \end{aligned} \tag{3.51}$$

From (3.50) and (3.51), we obtain (3.49).

The matrix-valued Gaussian distribution arises when sampling from a multivariate Gaussian population. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ be a random sample of size N from $\mathcal{N}_p(\mathbf{m}, \boldsymbol{\Sigma})$. Define the observation matrix as

$$\mathbf{X} = \left(\mathbf{x}_1, \ldots, \mathbf{x}_N\right) = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1N} \\ X_{21} & X_{22} & \cdots & X_{2N} \\ \vdots & \vdots & \cdots & \vdots \\ X_{p1} & X_{p2} & \cdots & X_{pN} \end{pmatrix}$$

then $\mathbf{X}^T \sim \mathcal{N}_{N,p}\left(\mathbf{e}\mathbf{m}^T, \mathbf{I}_N \otimes \boldsymbol{\Sigma}\right)$, where $\mathbf{e}(N \times 1) = (1, \ldots, 1)^T$.

If $\mathbf{X} \sim \mathcal{N}_{n,p}(\mathbf{M}, \boldsymbol{\Sigma} \otimes \boldsymbol{\Psi})$, then $\mathbf{X}^T \sim \mathcal{N}_{n,p}\left(\mathbf{M}^T, \boldsymbol{\Psi} \otimes \boldsymbol{\Sigma}\right)$.

If $\mathbf{X} \sim \mathcal{N}_{n,p}(\mathbf{M}, \mathbf{\Sigma} \otimes \mathbf{\Psi})$, then the characteristic function of $\mathbf{X}$ is

$$\Phi_{\mathbf{X}}(\mathbf{Z}) = \exp\left(j\mathbf{Z}^T\mathbf{M} - \frac{1}{2}\mathbf{Z}^T\mathbf{\Sigma}\mathbf{Z}\mathbf{\Psi}\right) \tag{3.52}$$

Let us derive (3.52):

$$\begin{aligned}\Phi_{\mathbf{X}}(\mathbf{Z}) &= \mathbb{E}\left\{\exp\left(\mathrm{Tr}\left(j\mathbf{Z}^T\mathbf{Z}\right)\right)\right\} \\ &= \mathbb{E}\left\{\exp\left(j\,\mathrm{Tr}\left(\left(\mathrm{vec}\left(\mathbf{X}^T\right)\right)^T \mathrm{vec}\left(\mathbf{Z}^T\right)\right)\right)\right\}\end{aligned}$$

Now we know that $\mathrm{vec}\left(\mathbf{X}^T\right) \sim \mathcal{N}_{pn}\left(\mathrm{vec}\left(\mathbf{M}^T\right), \mathbf{\Sigma} \otimes \mathbf{\Psi}\right)$. Thus, from the characteristic function of a vector-valued Gaussian distribution, we obtain

$$\begin{aligned}\Phi_{\mathbf{X}}(\mathbf{Z}) &= \exp\left(j\left(\mathrm{vec}\left(\mathbf{X}^T\right)\right)^T \mathrm{vec}\left(\mathbf{Z}^T\right) - \tfrac{1}{2}\left(\mathrm{vec}\left(\mathbf{Z}^T\right)\right)^T (\mathbf{\Sigma} \otimes \mathbf{\Psi})\,\mathrm{vec}\left(\mathbf{Z}^T\right)\right) \\ &= \exp\left(\mathrm{Tr}\left(j\mathbf{Z}^T\mathbf{M} - \tfrac{1}{2}\mathbf{Z}\mathbf{\Sigma}\mathbf{Z}\mathbf{\Psi}\right)\right)\end{aligned}$$

The last equality follows from Theorem 3.10.2.

3.11 Matrix-Valued Wishart Distribution

See Section B.3

3.12 Moment Method

Figure 3.2 gives the intuition for finding statistical metrics for hypothesis tests because the closed-form expression agrees with empirical simulations. We must take advantage of the closed-form expression offered by the Marchenko–Pastur law.

Consider the motivated hypothesis problem of (8.2). For hypothesis $\mathcal{H}_0$, we deal with $\mathbf{X}\mathbf{X}^H$. If we assume that conditions for random matrix $\mathbf{X}$ are met for Theorem 3.6.1, we can apply the Marchenko–Pastur law for hypothesis $\mathcal{H}_0$.

On the other hand, for hypothesis $\mathcal{H}_1$, we deal with

$$\mathbf{Y}\mathbf{Y}^H = \mathrm{SNR} \cdot \mathbf{H}\mathbf{H}^H + \mathbf{X}\mathbf{X}^H + \sqrt{\mathrm{SNR}}\left(\mathbf{H}\mathbf{X}^H + \mathbf{X}\mathbf{H}^H\right)$$

which is different from $\mathcal{H}_0$. Our intuition says that the additional terms in the above expression of $\mathbf{Y}\mathbf{Y}^H$ will *deform* the resultant distribution such that the distribution of $\mathbf{Y}\mathbf{Y}^H$ deviates from that of $\mathbf{X}\mathbf{X}^H$, whose distribution of eigenvalues follows the Marchenko–Pastur distribtuion. To take advantage of this deformation, our task is to find statistical metrics to measure it. It seems natural to choose the moments of the empirical spectral density as such metrics.

Similar to the proof of the semicircle law, we use the trace relation: for a positive integer k, the k-th moment of the empirical spectral density is given by

$$m_k = \int x^k F_{\mathbf{S}}(dx) = \frac{1}{N}\mathrm{Tr}\left(\mathbf{S}^k\right) = \frac{1}{n}\mathrm{Tr}\left(\left(\frac{1}{n}\mathbf{X}^H\mathbf{X}\right)^k\right) \tag{3.53}$$

For the Marchenko–Pastur distribution, the moments are given by

$$m_{k,MP} = \int_a^b x^k \rho_{MP}(x)dx = \sum_{i=0}^{k-1} \frac{1}{i+1}\binom{k}{i}\binom{k-1}{i} y^i \tag{3.54}$$

for $k \geq 0$. See [193] for the derivation of (3.54). When $k = 0$, the zero-order moment m_0 is the area under the curve of $\rho_{MP}(x)$, as illustrated in Figure 3.2.

The expectation of moments is

$$\mathbb{E}\left(m_k\right) = \mathbb{E}\left(\frac{1}{N}\operatorname{Tr}\left(\mathbf{S}^k\right)\right) = \mathbb{E}\left\{\frac{1}{n}\operatorname{Tr}\left(\left(\frac{1}{n}\mathbf{X}^H\mathbf{X}\right)^k\right)\right\} \tag{3.55}$$

For each fixed integer k, it follows from [163] that

$$\mathbb{E}\left(\frac{1}{N}\operatorname{Tr}\left(\mathbf{S}^k\right)\right) = \sum_{i=1}^{k-1}\left(\frac{N}{n}\right)^i \frac{1}{i+1}\binom{k}{i}\binom{k-1}{i} + O\left(\frac{1}{n}\right) \tag{3.56}$$

and

$$\operatorname{Var}\left(\frac{1}{N}\operatorname{Tr}\left(\mathbf{S}^k\right)\right) = O\left(\frac{1}{n^2}\right) \tag{3.57}$$

3.13 Stieltjes Transform Method

The Stieltjes transform is another fundamental tool, in addition to the moment method. The Stieltjes transform $s(z)$ of a Hermitian matrix $\mathbf{A}$ of $n \times n$ is defined for any complex number z not in the support of $F_{\mathbf{A}}(x)$:

$$s(z) = \int_{\mathbb{R}} \frac{1}{x-z} dF_{\mathbf{A}}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\lambda_i(\mathbf{A}) - z} \tag{3.58}$$

The Stieltjes transform can be regarded as the generating function of the moments from the observations: for z large enough

$$s(z) = \frac{1}{n}\operatorname{Tr}\left(\mathbf{A} - z\mathbf{I}\right)^{-1} = -\frac{1}{n}\sum_{k=0}^{n}\frac{1}{z^{k+1}}\operatorname{Tr}\left(\mathbf{A}^k\right) = -\frac{1}{n}\sum_{k=0}^{n}\frac{1}{z^{k+1}}m_k$$

The Stieltjes transform of the Marchenko–Pastur distribution is given by

$$s_{MP}(z) = \int_{\mathbb{R}} \frac{1}{x-z}\rho_{MP}(x)dx = \int_a^b \frac{1}{2\pi xy}\sqrt{(b-x)(x-a)}dx$$

which is the unique solution to the equation

$$s_{MP}(z) + \frac{1}{y+z-1+yzs_{MP}(z)} = 0$$

in the upper plane.

Some manipulations give

$$s_{MP}(z) = -\frac{y+z-1-\sqrt{\left(y+z-1\right)^2 - 4yz}}{2yz}$$

where we take the branch of $\sqrt{\left(y+z-1\right)^2 - 4yz}$ with cut at $[a, b]$ that is asymptotically $y + z - 1$ as $z \to \infty$.

Proposition 3.13.1 (criterion of convergence—Section 2.4 in [67]) Let μ_n be a sequence of probability measure defined on the real line and μ be a deterministic probability measure. Then μ_n converges to μ in probability if and only if $s_{\mu_n}(z)$ converges to $s_\mu(z)$ in probability for every z in the upper half plane.

The notion of convergence in probability is defined in (3.9). By Proposition 3.13.1, the Marchenko–Pastur law follows from the criterion of convergence by showing that

$$s(z) \to s_{MP}(z)$$

in probability for every z in the upper half plane.

A more careful analysis of the Stieltjes transform $s(z)$ gives more accurate and powerful control of the empirical spectral density of $\mathbf{A}$. We are interested in the local version of the Marchenko–Pastur law $\rho_{MP}(x)$, which is defined in (3.10).

Consider the sample covariance matrix $\mathbf{S} = \frac{1}{n}\mathbf{X}^H\mathbf{X}$, where $\mathbf{X} = \left\{\xi_{ij}\right\}_{1\leqslant i\leqslant N, 1\leqslant j\leqslant n}$ is a rectangular random matrix with entries bounded by K where K may depend on n.

Let $N_I(\mathbf{A})$ denote the number of eigenvalues of $\mathbf{A}$ on the interval I. The length of an interval is denoted by $|I|$. As illustrated in Figure 3.2, it is natural to ask how many eigenvalues of $\mathbf{S}$ lie on the interval I if the length $|I|$ shrinks with n. This problem lies at the heart of proving universality of the local eigenvalue statistics: see [182, 194] and [195].

For any constants $\varepsilon, \delta, C_1 > 0$, there exists $C_2 > 0$ such that the following holds. Assume that $N/n \to y$ for some $0 < y \leq 1$. Then, with probability at least $1 - n^{-C_1}$, one has [193]

$$\left| N_I(\mathbf{S}) - N\int_I \rho_{MP}(x)dx \right| \leqslant \delta N\,|I| \tag{3.59}$$

for any interval $I \subset (a+\varepsilon, b-\varepsilon)$ of length $|I| \geq C_2K^2 \log n/n$.

3.14 Concentration of the Spectral Measure for Large Random Matrices

In general we do not know a priori that $\frac{1}{N}\operatorname{Tr} f\left(\mathbf{X}_A(\omega)\right)$ converges. The interest in Corollary 3.14.2 is in the case where N and M are large and N/M remains bounded and bounded away from zero.

Let the set $\mathcal{M}^H_{N\times N}$ be the set of complex entries $N \times N$ Hermitian matrices. Let f be a real valued function on $\mathbb{R}$. f can also be seen as a function from $\mathcal{M}^H_{N\times N}(\mathbb{C})$ into $\mathcal{M}^H_{N\times N}(\mathbb{C})$. For $\mathbf{M} \in \mathcal{M}^H_{N\times N}(\mathbb{C})$, and the eigenvalue decomposition $\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{U}^H$ for a diagonal real matrix $\mathbf{D}$ and a unitary matrix $\mathbf{U}$, we have

$$f(\mathbf{M}) = \mathbf{U}f(\mathbf{D})\,\mathbf{U}^H$$

where $f(\mathbf{D})$ is the diagonal matrix with entries $\left(f\left(D_{11}\right), \ldots, f\left(D_{NN}\right)\right)$

Often we are interested in $\frac{1}{N}\operatorname{Tr} f(\mathbf{M})$, for $\mathbf{M} \in \mathcal{M}^H_{N\times N}(\mathbb{C})$. We consider the concentration of the real valued random variable $\frac{1}{N}\operatorname{Tr} f\left(\mathbf{X}_A\right)$ for inhomogeneous random matrices given by

$$\mathbf{X}_A = \left(\left(\mathbf{X}_A\right)_{ij}\right)_{1\leqslant i,j\leqslant N}, \quad \mathbf{X}_A = \mathbf{X}_A^H, \quad \left(\mathbf{X}_A\right)_{ij} = \frac{1}{\sqrt{N}}A_{ij}\omega_{ij}$$

with

$$\boldsymbol{\omega} := \left(\omega^R + i\omega^I\right) = \left(\omega_{ij}\right)_{1\leqslant i,j\leqslant N} = \left(\omega_{ij}^R + \sqrt{-1}\omega_{ij}^I\right)_{1\leqslant i,j\leqslant N}, \quad \omega_{ij} = \overline{\omega}_{ji}$$

$$\mathbf{A} = \left(A_{ij}\right)_{1\leqslant i,j\leqslant N}, \quad A_{ij} = \overline{A}_{ji}$$

where $\left(\omega_{ij}\right)_{1\leqslant i,j\leqslant N}$ are independent complex random variables with laws $\left(P_{ij}\right)_{1\leqslant i\leqslant N, 1\leqslant j\leqslant M}$, P_{ij} being a probability measure on $\mathbb{C}$ with

$$P_{ij}\left(\omega_{ij} \in \bullet\right) = \int \mathbb{I}_{u+iv\in\bullet} P_{ij}^R(du) P_{ij}^I(dv)$$

and $\mathbf{A}$ is a nonrandom complex matrix with entries $\left(A_{ij}\right)_{1\leqslant i,j\leqslant N}$ uniformly bounded by, say, a.

When needed, we shall write $\mathbf{X}_A = \mathbf{X}_A(\boldsymbol{\omega})$. We let $\Omega_N = \left\{\left(\omega^R, \omega^I\right)\right\}_{1\leqslant i,j\leqslant N}$, and denote by $\mathbb{P}^N$ the law $\mathbb{P}^N = \otimes_{1\leqslant i\leqslant j\leqslant N}\left(P_{ij}^R \otimes P_{ij}^I\right)$ on Ω_N, with $P_{ii}^I = \delta_0$, where δ_0 is the Dirac measure.

For a compact set K, we denote its diameter by $|K|$, that is the maximal distance between two points of K. For a Lipschitz function $f : \mathbb{R}^n \mapsto \mathbb{R}$, we define the Lipschitz constant $|f|_{\mathcal{L}}$ by

$$|f|_{\mathcal{L}} = \sup_{\mathbf{x},\mathbf{y}} \frac{|f(\mathbf{x}) - f(\mathbf{y})|}{\|\mathbf{x} - \mathbf{y}\|}$$

where $||\cdot||$ denotes the Euclidean norm on $\mathbb{R}^n$.

We say that a measure ν on $\mathbb{R}$ satisfies the logarithmic Sobolev inequality with the (not necessarily optimal) constant c if, for any differentiable function f,

$$\int f^2 \log \frac{f^2}{\int f^2 d\nu} d\nu \leqslant 2c \int \left|f'\right|^2 d\nu$$

where $f'(x)$ is the first derivative of $f(x)$. Recall that a measure ν satisfying the logarithmic Sobolev inequality possesses sub-Gaussian tails. Recall also that the Gaussian law [196] that any probability measure ν absolutely continuous with respect to the Lebesgue measure satisfying the Bobkov and Götze [197] condition (including $\nu(dx) = Z^{-1}e^{-|x|^\alpha}dx$ for $\alpha \geq 2$), as well as any distribution absolutely continuous with respect to them possessing an upper and lower bounded density, satisfies the logarithmic Sobolev inequality [198, Section 7.1].

Theorem 3.14.1 ([199]) (a) Assume that the $\left(P_{ij}\right)_{i\leqslant j, i,j\in\mathbb{N}}$ are uniformly compactly supported, that is that there exists a compact set $K \subset \mathbb{C}$, so that for any $1 \leqslant i \leqslant j \leqslant N$, $P_{ij}(K^c) = 0$. Assume that $f(x)$ is convex and Lipschitz. Then, for any $t > t_0(N) := 8|K|\sqrt{\pi}a|f|_{\mathcal{L}}/N > 0$

$$\mathbb{P}^N\left(\left|\frac{1}{N}\operatorname{Tr} f\left(\mathbf{X}_A(\omega)\right) - \mathbb{E}\frac{1}{N}\operatorname{Tr} f\left(\mathbf{X}_A\right)\right| > t\right)$$
$$\leqslant 4\exp\left\{-\frac{N^2\left(t - t_0(N)\right)^2}{16|K|^2 a^2 |f|_{\mathcal{L}}^2}\right\}$$

(b) If the $\left(P_{ij}^R, P_{ij}^I\right)_{1\leqslant i\leqslant j\leqslant N}$ satisfy the logarithmic Sobolev inequality with uniform constant c, then for any Lipschitz function f, for any $t > 0$

$$\mathbb{P}^N\left(\left|\frac{1}{N}\operatorname{Tr} f\left(\mathbf{X}_A(\omega)\right) - \mathbb{E}\frac{1}{N}\operatorname{Tr} f\left(\mathbf{X}_A\right)\right| > t\right) \leqslant 2\exp\left\{-\frac{N^2t^2}{16ca^2|f|_{\mathcal{L}}^2}\right\}$$

The well known Wishart's matrices (or sample covariance matrices) are used throughout the book. We shall state results under the natural normalization

$$\int xP_{ij}(dx) = 0, \qquad \int x^2P_{ij}(dx) = 1$$

If $\mathbf{Y}$ is a $N\times M$ matrix, $N \leq M$, with independent entries $\omega_{ij} = \operatorname{Re}\left(\omega_{ij}\right) + i\operatorname{Im}\left(\omega_{ij}\right)$ of law P_{ij}, $\mathbf{Z} = \mathbf{Y}\mathbf{Y}^H$ is a so-called Wishart's matrix. Let $\mathbb{P}^{N,M} = \otimes_{1\leqslant i\leqslant N, 1\leqslant j\leqslant M} P_{ij}$. For the sake of completeness, let us consider inhomogeneous Wishart matrices given for a diagonal real matrix $\mathbf{R} = \left(\lambda_1, \ldots, \lambda_M\right)$ with $\lambda_i \geqslant 0$ by $\mathbf{Z} = \mathbf{Y}\mathbf{R}\mathbf{Y}^H$. To deduce the concentration of the spectral measure for such matrices from Theorem 3.14.1, note that if we consider

$$\begin{aligned}
A_{ij} &= 0 \quad \text{for } 1 \leqslant i \leqslant N, 1 \leqslant j \leqslant N\\
A_{ij} &= 0 \quad \text{for } M+1 \leqslant i \leqslant M+N, M+1 \leqslant j \leqslant M+N\\
A_{ij} &= \sqrt{\lambda_{i-M}} \quad \text{for N} + 1 \leqslant i \leqslant M+N, 1 \leqslant j \leqslant N\\
A_{ij} &= \sqrt{\lambda_{j-M}} \quad \text{for } 1 \leqslant i \leqslant N, N+1 \leqslant j \leqslant N+M
\end{aligned}$$

then $\mathbf{A} = \mathbf{A}^H$ and if we consider $\mathbf{X}_A \in \mathcal{M}^H_{(N+M)\times(N+M)}(\mathbb{C})$ constructed as in the previous section, $\mathbf{X}_A$ can be written as

$$\begin{pmatrix} 0 & \mathbf{Y}\mathbf{R}^{1/2} \\ \mathbf{R}^{1/2}\mathbf{Y} & 0 \end{pmatrix}$$

Now, it is straightforward to see that $\left(\mathbf{X}_A\right)^2$ is equal to

$$\begin{pmatrix} \mathbf{Y}\mathbf{R}\mathbf{Y}^H & 0 \\ 0 & \mathbf{R}^{1/2}\mathbf{Y}^H\mathbf{Y}\mathbf{R}^{1/2} \end{pmatrix}$$

In particular, for any measurable function f,

$$\operatorname{Tr} f\left(\left(\mathbf{X}_A\right)^2\right) = 2\operatorname{Tr} f\left(\mathbf{Y}\mathbf{R}\mathbf{Y}^H\right) + (M-N)f(0)$$

It is therefore a direct consequence of Theorem 3.14.1 that

Corollary 3.14.2 (Guionnet and Zeitouni (2000) [199]) With $\left(X_{ij}\right)_{1\leqslant i\leqslant N, 1\leqslant j\leqslant M}$ independent random variables, and with $\mathbb{P}^{N,M}$ defined above, we let $\mathbf{R}$ be a non-negative diagonal matrix with finite spectral radius ρ. Set $\mathbf{Z} = \mathbf{X}\mathbf{R}\mathbf{X}^H$. Then

- If the $\left(P_{ij}\right)_{1\leqslant i\leqslant N, 1\leqslant j\leqslant M}$ are supported in a compact set K, for any function f so that $g(x) = f(x^2)$ is convex and has finite Lipschitz norm $|g|_{\mathcal{L}} \equiv |||f|||_{\mathcal{L}}$, for any $t > t_0(N+M) := 4|K|\sqrt{\pi\rho}|||f|||_{\mathcal{L}}/(N+M)$

$$\mathbb{P}\left(\left|\frac{1}{N}\operatorname{Tr} f(\mathbf{Z}) - \mathbb{E}\frac{1}{N}\operatorname{Tr} f(\mathbf{Z})\right| > t\frac{M+N}{N}\right)$$
$$\leqslant 4\exp\left(-\frac{1}{4|K|^2\rho|||f|||_{\mathcal{L}}^2}\left(t - t_0(N+M)\right)^2(N+M)^2\right).$$

- If the $\left(P_{ij}\right)_{1\leqslant i\leqslant N, 1\leqslant j\leqslant M}$ satisfy the logarithmic Sobolev inequality with uniformly bounded constant c, the above result holds for any Lipschitz functions $g(x) = f(x^2)$: for any $t > 0$

$$\mathbb{P}\left(\left|\frac{1}{N}\operatorname{Tr} f(\mathbf{Z}) - \mathbb{E}\frac{1}{N}\operatorname{Tr} f(\mathbf{Z})\right| > t\frac{M+N}{N}\right)$$
$$\leqslant 2\exp\left(-\frac{1}{2c\rho|||f|||_{\mathcal{L}}^2}t^2(N+M)^2\right)$$

The proof is, in fact, straightforward because, with the above remarks, one should see $f(\mathbf{Z})$ as $g\left(\mathbf{X}_A\right) = f\left(\left(\mathbf{X}_A\right)^2\right)$ and thus control the Lipschitz norm of g and the convexity of g.

The results presented above extend to the case where $\mathbf{R}$ is self-adjoint non-negative but not diagonal.

3.15 Future Directions

The ubiquity of massive streaming data, especially in problems involving array signal processing, trade of stocks and various online trading schemes, and so forth, makes RMT ideal. Integration of computational tools for analysis of large dimensional data using RMT has the potential to create a new paradigm for statistical practices. Hadamard products of random matrices are applicable to missing at random scenarios.

The vision of studying the interactions among, RMT, big data, and smart grid is explicitly outlined in the previous books of the author [39,40], where RMT is used as the unifying theme. This book makes more ideas concrete. The connection of RMT with massive MIMO [200], viewed as an large array of antennas—in the order of $n = 800 - 1000$, may be fruitful. In some sense, massive MIMO is a problem of big data.

The data are dependent on time, and much of the theory in the field is under the setting of i.i.d. observations. The current setting of i.i.d. observations can be extended to the case when the columns of the data matrix can be viewed as a realization of a high-dimensional multivariate time series, $t = 1, \ldots, NT_s$, where T_s is the sampling interval. We deal with a data matrix $\mathbf{X}$ of size $n \times N$. The natural setting is when N and n can be arbitrary, including the case that $N/n \to c \in (0, \infty)$, as $n \to \infty, N \to \infty$.

Bibliographical Remarks

Some material from Section 3.1 can be found in [201]. The example on $T_n = \log\left(\det \mathbf{S}_n\right)$ is inspired by the introduction to the Bai and Silverstein (2010) [163].

Historically, the author's interest in random matrix theory was initially led by the work in [61], where it was found that the classical methods such as generalized likelihood ratio test (GLRT) were outperformed by the generalized functions of sample covariance matrices. In other words, we proposed the new statistic $f\left(\mathbf{S}_n\right)$, where $f(x)$ was some very general function. With research going deeper and deeper, we understood that the $\mathbf{S}_n$ might be viewed as a random matrix that is also a non-negative, Hermitian matrix. The most critical step was made when we realized that the trace function gave the best empirical results in MATLAB simulations. Then we were convinced that $\mathrm{Tr}\, f\left(\mathbf{S}_n\right)$ was the new statistic for our problem.

Paul and Aue (2013) [38] give an overview of random matrix theory (RMT) with the objective of highlighting the results and concepts that have a growing impact in the formulation and inference of statistical models and methodologies, in the context of high-dimensional statistics. We drew on some material from this paper in Section 3.15.

We took some material from [202–204] in Section 3.2. Some good tutorials on random matrix theory are [164,183], where some MATLAB codes are available. The material in Section 3.3, Section 3.4 and Section 3.5 is taken from [178].

We drew material freely from the Ph.D. dissertation of Wang [193], in particular in Section 3.6. Most of results are standard in the literature.

Classical works on covariance matrices [172, 180–182] are still inspiring. The goal of [205] is to prove the central limit theorem for linear statistics of the eigenvalues of real symmetric band random matrices with independent entries.

In [206], the authors study a Wigner matrix $\mathbf{H}$—a random $N \times N$ matrix whose entries are independent up to symmetry constraints—which has been deformed by the addition of a finite-rank matrix $\mathbf{A}$ belonging to the same symmetry class as $\mathbf{H}$. By Weyl's eigenvalue interlacing inequalities, such a deformation does not influence the global statistics of the eigenvalues as $N \to \infty$.

According to [52], there is a long history for the celebrated log-det formula

$$\log\det\left(\mathbf{I} + \mathrm{SNR}\,\mathbf{H}\mathbf{H}^H\right)$$

where $\mathbf{H}$ is a random matrix. In 1964, Pinsker [207] gave a general log-det formula for the mutual information between jointly Gaussian random vectors but did not specifically work on the linear model (3.11). Verdu [208], in 1986, gave the explicit form of (3.12) as the capacity of the synchronous DS-CDMA channel as a function of signature vectors. The 1991 edition of the textbook by Cover and Thomas [59] gives the log-det formula for the capacity of the power constrained vector Gaussian channel with arbitrary noise covariance matrix. In the mid 1990s, Foschini [209] and Telatar [210] gave (3.12) for the multiple-input, multiple-output (MIMO) channel with i.i.d. Gaussian entries. The analysis of Gaussian channel with memory via vector channels (e.g. [211,212]) used the fact that the capacity can be expressed as the sum of the capacities of independent channels whose signal-to-noise ratios are governed by the singular values of the channel matrix. Recently, authors in [213] applied log-determinants of random matrices in the context of information theory, exploiting concentration inequalities for random matrices.

Example 3.6.3 is taken from [54].

Central limit theorem for the linear eigenvalue statistics of the Wigner and sample covariance matrix ensemble was proved in [181]. In Section 3.7, we followed [214,215]. Analogous results for linear statistics of the eigenvalues of symmetric band random

matrices with independent entries are studied in [205]. In [216], they consider a class of real random matrices with dependent entries and show that the limiting empirical spectral distribution is given by the Marchenko–Pastur law. They also establish a rate of convergence for the expected empirical spectral distribution.

Section 3.14 is taken from [199].

Example 3.9.1 and Example 3.9.2 are adapted from Forrester (2010) [62]. We modified the results for two independent random matrices. Example 3.9.3 is taken from Bai and Silverstein (2010) [163]. Example 3.9.4 is adapted from Tao (2012) [67].

In Section 3.9, we take some definitions from [217] on random matrices.

Sections 3.10 and 3.11 are taken from [217].

4

Linear Spectral Statistics of the Sample Covariance Matrix

This chapter, by studying the central limit theory for linear spectral statistics, conducts the spectral analysis of large dimensional random matrices. It is a fundamental work because many important statistics in multivariate statistical analysis can be expressed as functionals of the empirical spectral distribution of some random matrices. Thus, a deeper investigation of the convergence of the empirical spectral distribution is needed for more efficient statistical inferences, such as tests of hypotheses and confidence regions.

4.1 Linear Spectral Statistics

One of the exciting developments in statistics since the mid-1990s has been the development of theory and methodologies for dealing with high-dimensional data. The term "dimension" is primarily interpreted as meaning that the dimensionality of the observed multivariate data is comparable to the available number of replicates or subjects on which the measurements on the different variables are taken. This is often expressed in the asymptotic framework as $n \to \infty$, $p \to \infty$, such that $p/n \to c > 0$, where p denotes the dimension of the observation vectors (forming a triangular array) and n denotes the sample size.

One very notable high-dimensional phenomenon associated with sample covariance matrices is that the sample eigenvalues do not converge to their population counterparts if dimension and sample sizes remain comparable even as the sample size increases. A formal way to express this phenomenon is through the use of the empirical spectral distribution (ESD), that is, the empirical distribution of the eigenvalues of the sample covariance matrix.

The empirical spectral distribution (ESD) of the sample covariance matrix almost certainly converges to a nonrandom probability distribution known as the Marchenko–Pastur distribution (law). Since this highly influential discovery a large body of literature under the banner of random matrix theory (RMT) has been developed to explore the properties of the eigenvalues and eigenvectors of large random matrices.

For example, let $\mathbf{A}$ be an $n \times n$ positive definite matrix. Then

$$\frac{1}{n}\ln\det(\mathbf{A}) = \frac{1}{n}\sum_{i=1}^{n}\ln\lambda_i = \int_0^\infty \ln x dF_{\mathbf{A}}(x)$$

Generalizing the above example, we have the definition of a linear spectral statistic.

Smart Grid using Big Data Analytics: A Random Matrix Theory Approach, First Edition.
Robert C. Qiu and Paul Antonik.

Definition 4.1.1 (linear spectral statistic (LSS)) Let $F_n(x)$ be the empirical spectral distribution of a random matrix that has a limiting spectral distribution $F(x)$. We call

$$\hat{\theta} = \int f(x) dF_n(x) = \frac{1}{n} \sum_{i=1}^{n} f\left(\lambda_i\right)$$

a **linear spectral statistic (LSS)**.

Associated with the given random matrix, a linear spectral statistic can be considered as an estimator of

$$\theta = \int f(x) dF(x)$$

To make a test hypotheses about θ, it is necessary to know the limiting distribution of

$$G_n(f) = \alpha_n\left(\hat{\theta} - \theta\right) = \int f(x) dX_n(x)$$

where $X_n(x) = \alpha_n\left(F_n(x) - F(x)\right)$ and $\alpha_n \to \infty$ is a suitable normalizer such that $G_n(f)$ tends to a nondegenerate distribution.

4.2 Generalized Marchenko–Pastur Distributions

In Section 3.5, the population covariance matrix has the simple form $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_p$, which is quite restrictive. In order to consider a general population covariance matrix $\mathbf{\Sigma}$, we assume the following: the observed vectors $\left\{\mathbf{y}_k\right\}_{1 \leqslant k \leqslant n}$ can be expressed as

$$\mathbf{y}_k = \mathbf{\Sigma}^{1/2} \mathbf{x}_k$$

where $\mathbf{x}_k$ have i.i.d. components as in Section 3.5 and $\mathbf{\Sigma}^{1/2}$ is **any** non-negative squared root of $\mathbf{\Sigma}$. The associated sample covariance matrix is

$$\mathbf{B}_n = \frac{1}{n} \sum_{k=1}^{n} \mathbf{y}_k \mathbf{y}_k^H = \mathbf{\Sigma}^{1/2} \left(\frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_k \mathbf{x}_k^H \right) \mathbf{\Sigma}^{1/2} = \mathbf{\Sigma}^{1/2} \mathbf{S}_n \mathbf{\Sigma}^{1/2}.$$

Here $\mathbf{S}_n$ denotes the sample covariance matrix with i.i.d. components. The eigenvalues of $\mathbf{B}_n$ are the same as the product $\mathbf{S}_n \mathbf{\Sigma}$, for **all** non-negative matrices $\mathbf{\Sigma}$.

Proposition 4.2.1 (Bai and Silverstein (2010)) Let $\mathbf{S}_n$ be the sample covariance matrix with i.i.d. components and $\left(\mathbf{\Sigma}_n\right)_{n \geqslant 1}$ be a sequence of non-negative Hermitian squared matrices of size p. Let $\mathbf{B}_n = \mathbf{S}_n \mathbf{\Sigma}_n$. We assume that:

- the coordinates of $\mathbf{x}_i$ are complex i.i.d. with mean zero and variance one;
- the ratio of the data dimension over the sample size $p/n \to c > 0$ as $n \to \infty$;
- the sequence $\left(\mathbf{\Sigma}_n\right)_{n \geqslant 0}$ is deterministic, or independent from $\left(\mathbf{S}_n\right)_{n \geqslant 1}$;
- the sequence $\left(H_n\right)_{n \geqslant 0} = \left(F_{\mathbf{\Sigma}_n}\right)_{n \geqslant 0}$ of the empirical spectral distributions of $\left(\mathbf{\Sigma}_n\right)_{n \geqslant 0}$ converges weakly to a fixed probability measure H. Then $F_{\mathbf{B}_n}(x)$ converges weakly

to a fixed probability measure $F_{c,H}(x)$, whose Stieltjes transform, denoted by $m(z)$, is implicitly defined by the equation

$$s(z) = \int \frac{1}{t(1-c-czs(z))} dH(t) \tag{4.1}$$

where $z \in \mathbb{C}^+$.

The implicit equation given above has an unique solution in the space of functions from $\mathbb{C}^+$ to $\mathbb{C}^+$. Moreover, the solution $s(z)$ of this equation has **no closed-form** expression, and this is the *unique* information that we know about the limiting spectral distribution $F_{c,H}(x)$.

There is another way to present the fundamental equation (4.1). Take the squared matrix of size n

$$\mathbf{A}_n = \frac{1}{n}\mathbf{X}^H \mathbf{\Sigma} \mathbf{X}$$

where $\mathbf{X}$ is defined as $\mathbf{X} = \left(\mathbf{x}_1, \ldots, \mathbf{x}_n\right) \in \mathbb{C}^{p \times n}$. The two matrices $\mathbf{A}$ and $\mathbf{B}$ have the same positive eigenvalues and their empirical spectral distributions satisfy

$$nF_{\mathbf{A}_n} - pF_{\mathbf{B}_n} = (n-p)\,\delta_0$$

Assuming that $p/n \to c > 0$, $F_{\mathbf{B}_n}$ has a limit $F_{c,H}^{\mathbf{B}}$ if, and only if, $F_{\mathbf{A}_n}$ has a limit$F_{c,H}^{\mathbf{A}}$. In this case, the limits satisfy

$$F_{c,H}^{\mathbf{A}} - F_{c,H}^{\mathbf{B}} = (1-c)\,\delta_0$$

and their respective Stieltjes transform $s_{\mathbf{A}}(z)$ and $s_{\mathbf{B}}(z)$ are linked to each other by

$$s_{\mathbf{A}}(z) = -\frac{1-c}{z} + cs_{\mathbf{B}}(z)$$

Replacing $s_{\mathbf{B}}(z)$ by $s_{\mathbf{A}}(z)$ in (4.1), we find

$$s_{\mathbf{A}}(z) = -\left(z - c\int \frac{t}{1+s_{\mathbf{A}}(z)} dH(t)\right)^{-1}$$

Then solving this equation with respect to z leads to

$$z = -\frac{1}{s_{\mathbf{A}}(z)} + c\int \frac{t}{1+s_{\mathbf{A}}(z)} dH(t) \tag{4.2}$$

which gives the inverse function of $s_{\mathbf{A}}(z)$. The equations (4.1) and (4.2) are of fundamental importance in the methods of statistical estimation, and are called "Marchenko–Pastur equations."

The limiting spectral distribution $F_{c,H}^{\mathbf{B}}$ and its companion $F_{c,H}^{\mathbf{A}}$ are called "generalized Marchenko–Pastur distributions" with indexes c and H. In the case where $\mathbf{\Sigma}_n = \mathbf{T}$, the limiting spectral distribution H of $\mathbf{T}$ is called "population spectral distribution."

4.2.1 Central Limit Theorem

In this subsection, we use $F_{c,H}(x)$ to denote $F_{c,H}^{\mathbf{B}}(x)$ to save notation.

In multivariate analysis, most of the population statistics can be written as a function of the empirical spectral distribution F_n of some random matrices:

$$\hat{\theta} = \int f(x)\, dF_n(x)$$

$\hat{\theta}$ is called a "linear spectral statistic," and can be considered as an estimator of

$$\theta = \int f(x)\, dF(x)$$

where F is the limiting spectral distribution of F_n.

If we consider the sample covariance matrix B_n, we saw in Section 4.2 that its empirical spectral distribution F_n converges weakly to a generalized Marchenko–Pastur distribution $F_{c,H}$. This consistency is not enough for a better statistical inference, for which a central limit theorem is often required. In this section, we will present the result of Bai and Silverstein (2004) [188].

We consider the following linear spectral statistic

$$\hat{\theta}(f) = \int f(x)\, dF_{\mathbf{B}_n}(x)$$

As the convergences $c_n \to c$ and $H_n \to H$ can be very slow, the difference

$$p\left(\hat{\theta}(f) - \int f(x)\, dF_{c,H}(x)\right)$$

could have no limit. As a result, we have to consider the limiting distribution of the normalized difference

$$p\left(\hat{\theta}(f) - \int f(x)\, dF_{c_n,H_n}(x)\right)$$

In the sequel, we will denote

$$G_n(x) = p\left(F_{\mathbf{B}_n}(x) - F_{c_n,H_n}(x)\right)$$

Proposition 4.2.2 We denote by (x_{jk}) the entries of the vector $\mathbf{x}_j$. We assume: (i) For all $\eta \geq 0$

$$\frac{1}{np}\sum_{j,k}\mathbb{E}\left(\left|x_{jk}\right|^4 \mathrm{I}\left(\left|x_{jk}\right| \geqslant \eta\sqrt{n}\right)\right) \to 0 \text{ as } n \to \infty$$

(ii) For all n, the $x_{ij} = x_{ij}^{(n)}, 1 \leqslant i \leqslant p, 0 \leqslant j \leqslant n$ are independent, and satisfy

$$\mathbb{E}\left|x_{ij}\right|^2 = 1, \quad \max_{i,j,n}\mathbb{E}\left|x_{jk}\right|^4 < \infty, \quad \frac{p}{n} \to y$$

(iii) $\mathbf{T}_n \in \mathbb{C}^{p\times p}$ is non-negative Hermitian, with a bounded spectral norm in p, and there is a cumulative distribution function H such that

$$H_n \equiv F_{\mathbf{T}_n} \xrightarrow{\mathcal{L}} H$$

Let $f_1, \ldots, f_k$ be analytic functions on an open set of $\mathbb{C}$, which includes the interval

$$\left[\liminf_n \lambda_{n,\min}^{\mathbf{T}_n}\mathbf{1}_{]0,1[}(y)\left(1-\sqrt{c}\right)^2, \limsup_n \lambda_{n,\max}^{\mathbf{T}_n}\left(1+\sqrt{c}\right)^2\right]$$

then

a) The random vectors $X_n(f_1), \ldots, X_n(f_k)$ are a tight sequence in n.

b) If x_{ij} and $\mathbf{T}_n$ are real, and $\mathbb{E}\left|x_{ij}\right|^4 = 3$, then

$$\left(X_n\left(f_1\right), \ldots, X_n\left(f_k\right)\right) \xrightarrow{\mathcal{L}} \left(X\left(f_1\right), \ldots, X\left(f_k\right)\right)$$

where $\left(X\left(f_1\right), \ldots, X\left(f_k\right)\right)$ is a k-dimensional Gaussian vector.

c) If x_{ij} is complex with $\mathbb{E}\left(x_{ij}\right)^2 = 0$ and $\mathbb{E}\left|x_{ij}\right|^4 = 2$, then (b) also holds, except the mean is zero and the covariance function is a half of the function given in (b).

Example 4.2.3 (corrected likelihood ratio test ([160])) Let $\mathbf{x} \in \mathbb{R}^p$ be a random vector such that

$$\mathbf{x} \sim \mathcal{N}\left(\mathbf{0}_p, \mathbf{\Sigma}_p\right)$$

We would like to test

$$\mathcal{H}_0 : \mathbf{\Sigma}_p = \mathbf{I}_p$$
$$\mathcal{H}_1 : \mathbf{\Sigma}_p \neq \mathbf{I}_p$$

If we want to test $\mathbf{\Sigma}_p = \mathbf{A}$, with a given $\mathbf{A} \in \mathbb{C}^{p\times p}$, we can go back to the null above by the transformation $\mathbf{y} = \mathbf{A}^{-1/2}\mathbf{x}$. And we work on $\mathbf{y}$ instead. Let $\left(\mathbf{x}_1, \ldots, \mathbf{x}_n\right)$ be an n-sample of $\mathbf{x}$ such that $p < n$ and $\mathbf{S}_n$ the sample covariance matrix. We define

$$K^\star = \operatorname{Tr} \mathbf{S}_n - \log\det \mathbf{S}_n - p \tag{4.3}$$

The likelihood ratio statistic is $K_n = nK^\star$. When p is fixed and $n \to \infty$

$$K_n \xrightarrow{\mathcal{L}} \chi^2_{\frac{1}{2}p(p+1)}$$

under $\mathcal{H}_0$. When p becomes large, however, K_n grows to infinity, which leads to a test with higher level than the given one. Thus it is necessary to construct a version of K_n suitable in large dimensional setting. Notice that

$$K^\star = \sum_{i=1}^{p} \left(\lambda_{n,i} - \ln \lambda_{n,i} - 1\right)$$

where $\left(\lambda_{n,i}\right)_{1\leqslant i\leqslant p}$ are the eigenvalues of $\mathbf{S}_n$. This is a linear spectral statistic. We will apply Proposition 4.2.2 to obtain the asymptotic distribution of K_n in large dimensional setting. By taking $\mathbf{T}_n = \mathbf{I}_p$, $\mathbf{B}_n$ becomes $\mathbf{S}_n$. Moreover, we have $H_n = H = F_{\mathbf{T}_n} = \delta_1$, and also $X_n\left(f\right) = \int_{\mathbb{R}} f(x) d\left(F_{\mathbf{S}_n} - F_{c_n}\right)(x)$.

Applying Proposition 4.2.2, we obtain the following result:

Proposition 4.2.4 We assume that the conditions in Proposition 4.2.2 hold. $K^\star$ is defined as in (4.3) and $g(x) = x - \ln x - 1$. Then, under $\mathcal{H}_0$ and when $n \to \infty$:

$$\tilde{K}_n = \frac{1}{\sqrt{v(c)}}\left(K^\star - p\int_{\mathbb{R}} g(x) dF_{c_n}(x) - m(c)\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

where $m(c) = -\frac{\log(1-c)}{2}$, and $v(c) = -2\log(1-c) - 2c$.

In large dimensions, the limiting distribution of K_n is not a χ^2 law any more, but a Gaussian law. For intermediate dimensions such as $p = 50$, the corrected LRT gives good results, whereas the traditional LRT performs poorly. □

4.2.2 Spiked Population Models

We consider observations of the form

$$\mathbf{x}_i = \mathbf{\Sigma}^{1/2}\mathbf{y}_i,$$

where $\mathbf{y}_i$ are i.i.d. vectors of size p, with mean 0, variance 1, and i.i.d. components. $\left(\mathbf{x}_i\right)_{i\geqslant 1}$ is thus a random sequence of i.i.d. vectors with mean zero and population covariance matrix $\mathbf{\Sigma}$. If we take $\mathbf{\Sigma} = \mathbf{I}_p$, then this corresponds to the "null" case, and we saw from above that the limiting spectral distribution of $\mathbf{S}_n$ is the standard Marchenko–Pastur law. As noticed in [177], there are examples of real data that are significantly different from this null case. The so-called "spiked population model" is defined by the population covariance matrix $\mathbf{\Sigma}$ whose eigenvalues have the form

$$\underbrace{\alpha_1, \dots, \alpha_1}_{n_1}, \dots, \underbrace{\alpha_k, \dots, \alpha_k}_{n_K}, \underbrace{1, \dots, 1}_{p-m} \tag{4.4}$$

where $n_1 + \cdots + n_K = m$ is the number of "spikes." The spiked population model can be viewed as a finite rank perturbation of the null case.

When $p/n \to c > 0$, it is easy to see that the empirical spectral distribution of $\mathbf{S}_n$ still converges to the standard Marchenko–Pastur law. However, the asymptotic behavior of the extreme eigenvalues of $\mathbf{S}_n$ will be different from the null case.

4.2.3 Generalized Spiked Population Model

Bai and Yao (2012) [218] generalized the above model to a "generalized spiked population model." We assume that $\mathbf{\Sigma}_p$ has the following structure

$$\mathbf{\Sigma}_p = \begin{pmatrix} \mathbf{V}_m & 0 \\ 0 & \mathbf{T}_{p-m} \end{pmatrix}$$

Moreover, we assume:

- $\mathbf{V}_m$ is squared matrix of size m, where m is a fixed integer. The eigenvalues of $\mathbf{V}_m$ are $\alpha_1 > \cdots > \alpha_K > 0$ with respective multiplicities $n_1, \dots, n_K$ ($n_1 + \cdots + n_K = m$).
- The empirical spectral distribution H_p of $\mathbf{T}_{p-m}$ converges to a limiting non-random distribution H;
- The sequence of the largest eigenvalues of $\mathbf{\Sigma}$ is bounded;
- The eigenvalues $\beta_{n,j}$ of $\mathbf{T}_{p-m}$ satisfy

$$\sup_j d\left(\beta_{n,j}, \Gamma_H\right) = \varepsilon_p \to 0,$$

where $d(x, A)$ is the distance from x to a set A and Γ_H is the support of H.

Definition 4.2.5 An eigenvalue α of $\mathbf{V}_m$ is called a "generalized spike," or simply a spike, if $\alpha \notin \Gamma_H$.

Consequently, the spectrum of the population covariance matrix $\mathbf{\Sigma}$ is composed of a main part, $\beta_{n,j}$, and a smaller part of m spiked eigenvalues that are well separated from the main part, in the form of Definition 4.2.5.

Limits of spiked eigenvalues are given in [218]. The authors proved a central limit theorem for the vectors of eigenvalues.

4.3 Estimation of Spectral Density Functions

Our aim here is to recover the population spectral distribution $H(x)$ (or $H_N(x)$) from the sample covariance matrix $\mathbf{S}_n$. This task has a central importance in several popular statistical methodologies like principal component analysis [177], Kalman filtering or independent component analysis, which all rely on an efficient estimation of some population covariance matrices.

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a sequence of i.i.d. zero-mean random vectors in $\mathbb{R}^N$ or $\mathbb{C}^N$ with a common population (or true) covariance matrix $\mathbf{\Sigma}_N$. When the population size N is not negligible with respect to the sample size n, modern random matrix theory indicates that the sample covariance matrix

$$\mathbf{S}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^H$$

does not approach true $\mathbf{\Sigma}_N$. For instance, in a simple case where $\mathbf{\Sigma}_p = \mathbf{I}_N$ (identity matrix), the eigenvalues of $\mathbf{S}_n$ will spread over an interval approximately equal to $\left(1 \mp \sqrt{N/n}\right)^2$ around the unique population eigenvalue 1 of $\mathbf{\Sigma}_N$. Therefore, classical statistical procedures based on an approximation of $\mathbf{\Sigma}_N$ by $\mathbf{S}_n$ become inconsistent in such high-dimensional data situations.

The spectral distribution $G_{\mathbf{A}}$ of an $N \times N$ Hermitian matrix (or real symmetric) $\mathbf{A}$ is the measure generated by its eigenvalues $\left\{\lambda_i(\mathbf{A})\right\}_{i=1}^N$:

$$G_{\mathbf{A}}(x) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\lambda_i(\mathbf{A})}(x), \quad x \in \mathbb{R}$$

where δ_b denotes the Dirac point measure at b. Let $\left\{\lambda_i\left(\mathbf{\Sigma}_N\right)\right\}_{i=1}^N$ be the N eigenvalues of the true (or population) covariance matrix $\mathbf{\Sigma}_N$. We are particularly interested in the following spectral distribution

$$H_N(x) = G_{\mathbf{\Sigma}_N}(x) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\lambda_i(\mathbf{\Sigma}_N)}(x)$$

Following the random matrix theory, both sizes N and n will grow to infinity. It is, then, natural to assume that $H_N(x)$ weakly converges to a limiting distribution $H(x)$ when $N \to \infty$. We refer this limiting spectral distribution $H_N(x)$ as the population spectral distribution of the observation model.

The main observation is that under reasonable assumptions, when both dimensions N and n become large at a proportional rate say c, almost, the (random) spectral distribution $G_{\mathbf{S}_n}(x)$ of the sample covariance matrix $\mathbf{S}_n$ will weakly converge to a deterministic distribution $F(x)$, called limiting spectral distribution. Naturally this limiting spectral distribution $F(x)$ depends on the population spectral distribution $H(x)$, but in general this relationship is complex and has *no explicit form*. The only exception is the case where all the population eigenvalues $\left\{\lambda_i\left(\mathbf{\Sigma}_N\right)\right\}_{i=1}^N$ are unit: $\mathbf{\Sigma}_N = \mathbf{I}_N$, or $H(x) = \delta_1(x)$; the limiting spectral distribution $F(x)$ is then explicit known to be the Marchenko–Pastur distribution with an explicit density function. For a general population spectral distribution $H(x)$, this relationship is expressed via implicit equations, (4.5) and (4.7).

A critical task here is to recover the population spectral distribution $H(x)$ (or $H_N(x)$) from the sample covariance matrix $\mathbf{S}_n$.

Let $\mathbf{A}^{1/2}$ stand for any Hermitian square root of a non-negative definite Hermitian matrix $\mathbf{A} \geq 0$. Our model assumptions are as follows.

- The sample and population sizes n and N both tend to infinity, and in such a way that $N/n \to c \in (0, \infty)$.
- There is a doubly infinite array of i.i.d., complex-value random variables $\left(w_{ij}\right), i, j \geqslant 1$ satisfying

$$\mathbb{E}\left(w_{11}\right) = 0, \quad \mathbb{E}\left(\left|w_{11}\right|^2\right) = 1$$

  such that, for each N, n, letting $\mathbf{W}_n = \left(w_{ij}\right)_{1 \leqslant i \leqslant N, 1 \leqslant j \leqslant n}$, the observation vectors can be represented as $\mathbf{x}_j = \mathbf{\Sigma}_N^{1/2} w_{\cdot j}$, where $w_{\cdot j} = \left(w_{ij}\right)_{1 \leqslant i \leqslant N}$ denotes the j-th column of $\mathbf{W}_n$.
- The spectrum density $H_N(x)$ of $\mathbf{\Sigma}_N$ weakly converges to a probability distribution $H(x)$ as $n \to \infty$.

These assumptions are classical conditions for the celebrated Marchenko–Pastur theorem [163, 175, 219]. Under these assumptions, almost certainly, as $n \to \infty$, the empirical spectrum density $F_n(x) = G_{\mathbf{S}_n}(x)$ of the sample covariance matrix $\mathbf{S}_n$ weakly converges to a (nonrandom) generalized Marchenko–Pastur distribution $F(x)$.

Unfortunately, except in the simplest case where $H(x) \equiv \delta_1(x)$, the limit spectral density $F(x)$ has no explicit form and it is characterized as follows. Let $m(z)$ denote the Stieltjes transform of $cF(x) + (1-c)\delta_0(x)$, which is a one-to-one map defined on the upper half complex plane $\mathbb{C}^+ = \{z \in \mathbb{C} : \operatorname{Im}(z) > 0\}$. This transform $m(z)$ satisfies the following fundamental Marchenko–Pastur equation (MP):

$$z = -\frac{1}{m(z)} + c \int \frac{t}{1 + t m(z)} dH(t), \qquad z \in \mathbb{C}^+ \tag{4.5}$$

The above MP equation excludes the real line from its domain of definition. Following [220], we fill this gap by an extension of the MP equation to the real line. The estimation method can be entirely based on this extension.

4.3.1 Estimation Method

The support of a distribution $G(x)$ is denoted by $\operatorname{supp}(G)$ and its complementary set by $\operatorname{supp}^c(G)$. As the empirical spectral density $F_n(x)$ is observed, we will use $m_n(z)$ to approximate $m(z)$ in the MP equation. Here $m_n(z)$ is the Stieltjes transform of $(N/n) F_n(x) + (1 - N/n)\delta_0(x)$. More precisely, for $u \in \mathbb{R}$, let

$$m_n(u) = -\frac{1 - N/n}{u} + \frac{1}{n} \sum_{i=1}^{N} \frac{1}{\lambda_i - u} \tag{4.6}$$

It is clear that the domain of $m_n(u)$ is $\operatorname{supp}^c\left(F_n(x)\right)$. Thus, $m_n(u)$s are well defined on Ω_{interior} for all large n, where Ω_{interior} is the interior of Ω with $\Omega_{\text{interior}} = \liminf_{n \to \infty} \operatorname{supp}^c\left(F_n(x)\right) \setminus \{0\}$.

Theorem 4.3.1 Assume that the model assumptions hold. Then

- for any $u \in \Omega_{\text{interior}}$, $m_n(z)$ converges to $m(z)$;
- for any $u \in \text{supp}^c(F(x))$, $m(u)$ is a solution to equation

$$u = -\frac{1}{m_n(u)} + c\int \frac{t}{1+tm(u)}dH(t) \tag{4.7}$$

- the solution is also unique in the set

$$\mathrm{B} = \left\{ m(u) \in \mathbb{R}\setminus\{0\} : \frac{du}{dm(u)} > 0, (-m(u))^{-1} \in \text{supp}^c(H(x)) \right\}$$

- for any non-empty open interval $(a,b) \in \mathrm{B}$, $H(x)$ is uniquely determined by $u(m(u)), m(u) \in (a,b)$.

As $(-\infty, 0) \subset \Omega_{\text{interior}} \subset \text{supp}^c\left(F_n(x)\right)$, there are infinitely many u-points such that $m_n(u)$ almost surely converges to $m(u)$. The MP equation (4.7) can be inverted in the following sense: the knowledge of $u(m)$ on any interval in B will *uniquely* determine the population spectral distribution $H(x)$. The estimation method will be built on this property.

Now we are in a position to describe the estimation method, using the above theorem. Let us consider the estimation problem in a parametric setting. Suppose $H(x) = H(\theta)$ is the limit of H_N with unknown parameter vector $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$. The procedure of the estimation of $H(x)$ includes three steps:

1) Choose a u-net $\{u_1, \ldots, u_q\}$ from Ω_{interior}, where u_i's are distinct and the size q is no less than p.
2) For each u_i, calculate $m_n(u_i)$ using (4.6) and plug the pair into the Marchenko–Pastur equation (4.7). Then we obtain q approximate equations

$$u_i \approx -\frac{1}{m_n(u_i)} + \frac{N}{n}\int \frac{tdH(t,\boldsymbol{\theta})}{1+tm_n(u_i)} := \hat{u}_i\left(m_{ni}, \boldsymbol{\theta}\right), \quad i = 1, \ldots, q$$

3) Find the least-squares solution of $\boldsymbol{\theta}$:

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} \sum_{i=1}^{q} \left(u_i - \hat{u}_i\left(m_{ni}, \theta\right)\right)^2$$

See Figure 4.1. A central issue here is the choice of the u-net $\{u_1, \ldots, u_m\}$. For example, When $H(x)$ has finite support, the upper and lower bounds of $\text{supp}(F)\setminus\{0\}$ can be estimated respectively by $\lambda_{\max} = \max\{\lambda_i\}$ and $\lambda_{\min} = \min\{\lambda_i : \lambda_i > 0\}$ where λ_i are sample eigenvalues. As a consequence, we design a primary set:

$$U = \begin{cases} (-10,0) \cup \left(0, 0.5\lambda_{\max}\right) \cup \left(5\lambda_{\max}, 10\lambda_{\max}\right) & \text{(discrete model, } N \neq n\text{)} \\ (-10,0) \cup \left(5\lambda_{\max}, 10\lambda_{\max}\right) & \text{(discrete model, } N = n\text{)} \\ (-10,0) & \text{(continuous model)}. \end{cases}$$

Next, we choose ℓ equally spaced u-points from each individual interval of U. This process is called adaptive choice of u-net. Here we set $\ell = 20$ for all cases considered in simulation. For example we take $\{-10 + 10t/2\ell, t = 1, \ldots, 20\}$ from the first interval.

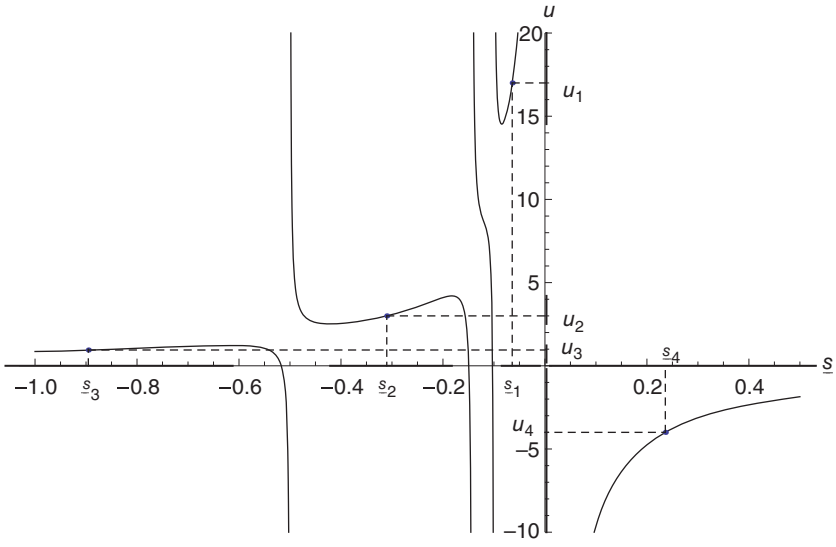


Figure 4.1 The curve of $u = u(m)$ (solid thin), and the sets B and $\text{supp}^c\left(F_n(x)\right)$ (solid thick) for $H(x) = 0.3\delta_2(x) + 0.4\delta_7(x) + 0.3\delta_{10}(x)$ and $c = 0.1$. $u_i = u\left(m_i\right), m_i \in \text{B}, i = 1, 2, 3, 4$. In the figure the $\underline{s}$ is the m in our notation. Source: Reproduced from [220] with permission.

4.3.2 Kernel Estimator of the Limiting Spectral Distribution

The density function of the limiting spectral distribution of general sample covariance matrices is usually unknown. We use kernel estimators, which have been proved to be consistent.

Suppose that X_{ij} are independent and identically distributed (i.i.d.) real random variables. Let $\mathbf{X}_n = \left(X_{ij}\right)_{N\times n}$ and $\mathbf{T}_N$ be an $N \times N$ nonrandom Hermitian non-negative definite matrix. Consider the random matrices

$$\mathbf{A}_n = \frac{1}{n}\mathbf{T}_N^{1/2}\mathbf{X}_n\mathbf{X}_n^T\mathbf{T}_N^{1/2}$$

When $\mathbb{E}X_{11} = 0$ and $\mathbb{E}X_{11}^2 = 1$, $\mathbf{A}_n$ can be viewed as a sample covariance matrix drawn from the population with true covariance matrix $\mathbf{T}_N$. Moreover, if $\mathbf{T}_N$ is another sample covariance matrix, independent of $\mathbf{X}_n$, then $\mathbf{A}_n$ is a Wishart matrix.

In order to capture the whole picture of the eigenvalues of sample covariance matrices, it is necessary to study the behavior of all eigenvalues. A good candidate for this purpose is the empirical spectral distribution. The basic limit theorem regarding $\mathbf{A}_n$ concerns its empirical spectral distribution $F_{\mathbf{A}_n}(x)$. Here, for any matrix $\mathbf{A}$ with real eigenvalues, the empirical spectral distribution $F_{\mathbf{A}}(x)$ is given by

$$F_{\mathbf{A}}(x) = \frac{1}{p}\sum_{i=1}^{p}\text{I}\left(\lambda_i(\mathbf{A}) \leqslant x\right)$$

where $\mathbb{I}$ is the indicator function, and $\lambda_i, i = 1, \ldots, p$, denote the eigenvalues of $\mathbf{A}$. Suppose the ratio of the dimension to the sample size $c_n = N/n$ tends to c as $n \to \infty$. When

$\mathbf{T}_n$ becomes the identity matrix, $\mathbf{T}_n = \mathbf{I}_n$, the so-called Marchenko–Pastur law with the density function

$$f_c(x) = \begin{cases} \frac{1}{2\pi c}\frac{1}{x}\sqrt{(b-x)(x-a)}, & a \leqslant x \leqslant b, \\ 0, & otherwise \end{cases} \tag{4.8}$$

It has point mass $1 - c^{-1}$ at the origin if $c > 1$, where $a = \left(1 - \sqrt{c}\right)^2$ and $b = \left(1 + \sqrt{c}\right)^2$. The distribution function of the Marchenko–Pastur law is denoted by $\mathbb{F}_c(x)$. The Stieltjes transform of the MP law is

$$m(z) = \frac{1 - c - z + \sqrt{(z-1-c)^2 - 4c}}{2cz} \tag{4.9}$$

which satisfies the equation

$$m(z) = \frac{1}{1 - c - czm(z) - z} \tag{4.10}$$

Here, the Stieltjes transform $m_F(z)$ for any probability distribution function $F(x)$ is defined by

$$m_F(z) = \int \frac{1}{x - z} dF(x), \qquad z \in \mathbb{C}^+ \tag{4.11}$$

where $\mathbb{C}^+ = \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$.

It is also common to study

$$\mathbf{B}_n = \frac{1}{n}\mathbf{X}_n^T \mathbf{T}_n \mathbf{X}_n$$

since the eigenvalues of $\mathbf{A}_n$ and $\mathbf{B}_n$ differ by $|n - p|$ zero eigenvalues. Thus

$$F_{\mathbf{B}_n}(x) = \left(1 - \frac{N}{n}\right) \mathrm{I}(x \in [0, \infty)) + \frac{N}{n} F_{\mathbf{A}_n}(x) \tag{4.12}$$

When $F_{\mathbf{T}_n}(x)$ converges weakly to a nonrandom distribution $H(x)$, Marchenko and Pastur (1967) [172], Yin (1986) [221] and Silverstein (1995) [175] proved that, with a probability of 1, $F_{\mathbf{B}_n}(x)$ converges in distribution to a nonrandom distribution function $F_{c,H}(x)$ whose Stieltjes transform $m_{F_{c,H}}(z)$ is, for each $z \in \mathbb{C}^+$, the unique solution to the fundamental equation of Marchenko and Pastur (4.5).

From (4.12), we have

$$G_{c,H}(x) = (1 - c)\,\mathrm{I}(x \in [0, \infty)) + cF_{c,H}(x)$$

where $F_{c,H}(x)$ is the limit of $F_{\mathbf{A}_n}(x)$. As a consequence of this fact, we have

$$m_G(z) = -\frac{1 - c}{z} + cm_F(z) \tag{4.13}$$

Moreover, $m_G(z)$ has an inverse

$$z\left(m_G(z)\right) = -\frac{1}{m_G(z)} + c_n \int \frac{t}{1 + tm_G(z)} dH(t) \tag{4.14}$$

Relying on this inverse, Silverstein and Choi [222] carried out a remarkable analysis of the analytic behavior of $m_G(z)$.

When $\mathbf{T}_n$ becomes the identity matrix, i.e., $\mathbf{T}_n = \mathbf{I}_n$, there is an explicit solution to the fundamental equation of Marchenko and Pastur (4.5). In this case, from (4.12), we see that the density function of $G_{c,H}(x)$ is

$$g_c(x) = (1-c)\,\mathrm{I}(c<1)\,\delta_0(x) + cf_c(x)$$

where $\delta_0(x)$ is the point mass at 0. Unfortunately, there is no explicit solution to (4.5) for general $\mathbf{T}_n$. Although we can use $F_{\mathbf{A}_n}(x)$ to estimate $F_{c,H}(x)$, we cannot make any statistical inference about $F_{c,H}(x)$ because there is no central limit theorem concerning $F_{\mathbf{A}_n}(x) - F_{c,H}(x)$. Actually, it is argued in [163] that the process $n\left(F_{\mathbf{A}_n}(x) - F_{c,H}(x)\right)$, for $x \in (-\infty, \infty)$, does not converge to a nontrivial process in *any* metric space. This motivates us to pursue alternative ways of understanding the limiting spectral distribution $F_{c,H}(x)$. Our aim is to estimate the density function $f_{c,H}(x)$ of the limiting spectral distribution $F_{c,H}(x)$ of sample covariance matrices $\mathbf{A}_n$ by kernel estimators.

Suppose that the observations $X_1, \ldots, X_n$ are i.i.d. random variables with an unknown density function $f(x)$ and $F_n(x)$ is the empirical distribution function determined by the sample. A popular nonparametric estimate of $f(x)$ is then

$$\hat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) = \frac{1}{h}\int K\left(\frac{x-y}{h}\right)dF_n(y) \tag{4.15}$$

where the function $K(y)$ is a Borel function and $h = h(n)$ is the bandwidth, which tends to 0 as $n \to \infty$. Here, $\hat{f}_n(x)$ is a probability density function and it inherits some smooth properties of $K(y)$ if the kernel is taken as a probability density function. Under some regularity conditions on the kernel, it is well known that $\hat{f}_n(x) \to f(x)$ in some sense (with probability 1). There is a huge body of literature regarding this kind of estimate. For example, we may refer to Rosenblatt [223], Parzen [224], Hall [225] or the books by Silverman [226] and Devroye and Lugosi [227].

With the aid of (4.15), we can consider the following estimator $f_n(x)$ of $f_{c,H}(x)$:

$$f_n(x) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x - \lambda_i\left(\mathbf{A}_n\right)}{h}\right) = \frac{1}{h}\int K\left(\frac{x-y}{h}\right)dF_{\mathbf{A}_n}(y) \tag{4.16}$$

where $\lambda_i\left(\mathbf{A}_n\right), i = 1, \ldots, n$, are eigenvalues of $\mathbf{A}_n$. It turns out that $f_n(x)$ is a *consistent* estimator of $f_{c,H}(x)$ under some regularity conditions. In (4.16), note the "smoothing" ideas. It was proved by [228] that $f_n(x)$ is a consistent estimator of $f_{c,H}(x)$ under some regularity conditions. The main aim of [229] is to establish a central limit theorem for this $f_n(x)$. This provides an approach to making inferences on the Marchenko–Pastur-type distribution functions.

The kernel estimator of the distribution function of the Marchenko–Pastur law is

$$F_n(x) = \int_{-\infty}^{x} f_n(y)\,dy$$

Intuitively, $F_n(x)$ depicts the global picture of all eigenvalues and should not differ greatly from $F_{\mathbf{A}_n}(x)$.

In the following, let us consider another example for the estimation of density functions. To do that, we first present the famous McDiarmid's concentration inequality.

Example 4.3.2 (concentration inequalities [230]) Hoeffding's inequality applies to sums of independent random variables. Its generalization to arbitrary real-valued functions of independent random variables that satisfy a certain condition is obtained in [231]. Let $\mathcal{X}$ be some set, and consider a function $g : \mathcal{X}^n \to \mathbb{R}$. We say that the function g has *bounded differences* if non-negative numbers exist $c_1, \ldots, c_n$, such that

$$\sup_{x_1,\ldots,x_n,x_i' \in \mathcal{X}} \left| g\left(x_1, \ldots, x_n\right) - g\left(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n\right) \right| \leqslant c_i \tag{4.17}$$

for all $i = 1, \ldots, n$. In words, if we change the i-th variable while keeping all the others fixed, the value of g will not change by more than c_i.

Theorem 4.3.3 (McDiarmid's inequality [231]) Let $X^n = \left(X_1, \ldots, X_n\right) \in \mathcal{X}^n$ be an n-tuple of independent $\mathcal{X}$-valued random variables. If a function $g : \mathcal{X}^n \to \mathbb{R}$ has bounded differences, as in (4.17), then, for all $t > 0$

$$\mathbb{P}\left(\left|g\left(X^n\right) - \mathbb{E}g\left(X^n\right)\right| \geqslant t\right) \leqslant \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}\right)$$

Let $X^n = \left(X_1, \ldots, X_n\right)$ be an n-tuple of i.i.d. real-valued random variables whose common distribution P has a probability density function (pdf) f:

$$P(A) = \int_A f(x)dx$$

for any measurable set $A \subseteq \mathbb{R}$. We wish to estimate f from the sample X^n. A popular method is to use a kernel estimate (the book by Devroye and Lugosi [227] has plenty of material on density estimation, including kernel methods, from the viewpoint of statistical learning theory). We pick a non-negative function $K(x) : \mathbb{R} \to \mathbb{R}$ that integrates to one, $\int K(x)dx = 1$ (such a function is called a *kernel*), as well as a positive bandwidth (or smoothing constant) $h > 0$ and form the estimate

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

$\hat{f}_n(x)$ is a valid probability distribution function, i.e., that it is non-negative and integrates to 1. A common way of quantifying the performance of a density estimator is to use the L_1 distance to the true density $f(x)$:

$$\left\|\hat{f}_n - f\right\|_{L_1} = \int_{\mathbb{R}} \left|\hat{f}_n(x) - f(x)\right| dx.$$

Obviously $\left\|\hat{f}_n - f\right\|_{L_1}$ is a random variable because it depends on the random sample X^n. Thus, we can write it as a function $g(X^n)$ of the sample X^n. Leaving aside the problem of actually bounding $\mathbb{E}g\left(X^n\right)$, we can easily establish a concentration bound for it using McDiarmid's inequality. To do that, we need to check that g has bounded differences. Choosing x^n and $x_{(i)}^n$, we have

$$
\begin{aligned}
& g\left(x^n\right)-g\left(x_{(i)}^n\right) \\
& =\int_{\mathbb{R}}\left|\frac{1}{nh}\sum_{i=1}^{j-1}K\left(\frac{x-x_i}{h}\right)+\frac{1}{nh}K\left(\frac{x-x_j^{'}}{h}\right)+\frac{1}{nh}\sum_{i=j+1}^{n}K\left(\frac{x-x_i}{h}\right)-f(x)\right|dx \\
& \quad -\int_{\mathbb{R}}\left|\frac{1}{nh}\sum_{i=1}^{j-1}K\left(\frac{x-x_i}{h}\right)+\frac{1}{nh}K\left(\frac{x-x_j^{'}}{h}\right)+\frac{1}{nh}\sum_{i=j+1}^{n}K\left(\frac{x-x_i}{h}\right)-f(x)\right|dx \\
& \leqslant \frac{1}{nh}\int_{\mathbb{R}}\left|K\left(\frac{x-x_j}{h}\right)-K\left(\frac{x-x_j^{'}}{h}\right)\right|dx \\
& \leqslant \frac{1}{nh}\int_{\mathbb{R}}K\left(\frac{x}{h}\right)dx=\frac{2}{n}
\end{aligned}
$$

Thus, we see that $g(X^n)$ has the bounded differences property with $c_1=\cdots=c_n=2/n$, so that

$$\mathbb{P}\left(\left|g\left(X^n\right)-\mathbb{E}g\left(X^n\right)\right|\geqslant t\right)\leqslant 2e^{-nt^2/2} \qquad \square$$

We are ready to present the main result in this section. Suppose that the kernel function $K(x)$ satisfies

$$\sup_{-\infty<x<\infty}|K(x)|<\infty, \qquad \lim_{|x|\to\infty}|xK(x)|=0 \tag{4.18}$$

and

$$\int K(x)dx=1, \qquad \int\left|K'(x)\right|dx<\infty \tag{4.19}$$

Theorem 4.3.4 Suppose that $K(x)$ satisfies (4.18) and (4.19). Let $h=h(n)$ be a sequence of positive constants satisfying

$$\lim_{n\to\infty}nh^{5/2}=\infty, \qquad \lim_{n\to\infty}h=0 \tag{4.20}$$

Moreover, suppose that all X_{ij} are i.i.d. with $\mathbb{E}X_{11}=0$, $\mathrm{Var}\left(X_{11}\right)=1$ and $\mathbb{E}X_{11}^{16}<\infty$. Also, assume that $c_n\to c\in(0,1)$. Let $\mathbf{T}_n$ be an $N\times N$ nonrandom symmetric positive definite matrix with spectral norm bounded above by a positive constant such that $H_n(x)=F_{\mathbf{T}_n}(x)$ converges weakly to a nonrandom distribution $H(x)$. In addition, suppose that $F_{c,H}(x)$ has a compact support $[a,b]$ with $a>0$. Then,

$$f_n(x)\to f_{c,H}(x) \qquad \textit{in probability uniformly in } x\in[a,b]$$

It is conjectured that $\mathbb{E}X_{11}^{16}$ can be reduced to $\mathbb{E}X_{11}^4<\infty$. When $\mathbf{T}_n$ is the identity matrix, we have a slightly better result.

Theorem 4.3.5 Suppose that $K(x)$ satisfies (4.18) and (4.19). Let $h=h(n)$ be a sequence of positive constants satisfying

$$\lim_{n\to\infty}nh^2=\infty, \qquad \lim_{n\to\infty}h=0 \tag{4.21}$$

Moreover, suppose that all X_{ij} are i.i.d. with $\mathbb{E}X_{11} = 0$, $\mathrm{Var}\left(X_{11}\right) = 1$ and $\mathbb{E}X_{11}^{12} < \infty$. Also, assume that $c_n \to c \in (0,1)$. Denote the support of the Marchenko–Pastur law by $[a,b]$. Let $\mathbf{T}_n = \mathbf{I}_n$. Then

$$\sup_{x\in[a,b]} \left|f_n(x) - f_c(x)\right| \to 0 \qquad \text{in probability}$$

Theorem 4.3.4 also leads to the estimate of $F_{c,H}(x)$, as below. Under the assumptions of Theorem 4.3.4, correspondingly,

$$F_n(x) \to F_{c,H}(x) \qquad \text{in probability} \tag{4.22}$$

where

$$F_n(x) = \int_{-\infty}^{x} f_n(t)dt \tag{4.23}$$

Using (4.22) and the Helly–Bray lemma, ensure that, under the assumptions of Theorem 4.3.4, if $g(x)$ is a continuous bounded function, then

$$\int g(x)dF_n(x) \to \int g(x)dF_{c,H}(x) \qquad \text{in probability} \tag{4.24}$$

In order to prove consistency of the nonparametric estimates, we need to develop a convergence rate for $F_{\mathbf{A}_n}(x)$. Under the assumptions of Theorem 4.3.4, we have

$$\sup_x \left|\mathbb{E}F_{\mathbf{A}_n}(x) - F_{c_n,H_n}(x)\right| = O\left(\frac{1}{n^{2/5}}\right) \tag{4.25}$$

and

$$\mathbb{E}\sup_x \left|F_{\mathbf{A}_n}(x) - F_{c_n,H_n}(x)\right| = O\left(\frac{1}{n^{2/5}}\right) \tag{4.26}$$

Under the fourth moment condition, that is, $\mathbb{E}X_{11}^4 < \infty$, it was proved in [229] that the above rate $O\left(n^{-2/5}\right)$ could be improved to $O\left(n^{-1}\sqrt{\log n}\right)$.

Example 4.3.6 (multiuser wireless systems) Since $F_{c,H}(x)$ does not have an explicit expression (except for some special cases), we may now use $F_n(x)$ to estimate it, by (4.22). More importantly, $F_n(x)$ has some *smoothness* properties, which $F_{\mathbf{A}_n}(x)$ does not have.

Consider a synchronous CDMA system with n users and processing gain N. The discrete time model for the received signal $\mathbf{Y}$ is given by

$$\mathbf{Y} = \sum_{k=1}^{n} x_k \mathbf{h}_k + \mathbf{W}$$

where $x_k \in \mathbb{R}$, and $\mathbf{h}_k \in \mathbb{R}^N$ are, respectively, the transmitted symbol and the signature spreading sequence of user k, and $\mathbf{W}$ is the Gaussian noise with zero mean and covariance matrix $\sigma^2\mathbf{I}_n$. Assume that the transmitted symbols of different users are independent, with $\mathbb{E}x_k = 0$, and $\mathbb{E}\left|x_k\right|^2 = P_k$. This model is slightly more general than that in [232], where all of the users' powers P_k are assumed to be the same.

Following [232], consider the demodulation of user 1 and use the signal-to-interference ratio (SIR) as the performance measure of linear receivers. For user 1, the

optimal demodulator $\mathbf{c}_1$ that generates a soft decision $\mathbf{c}_1^T\mathbf{Y}$ maximizing the signal-to-interference ratio

$$\beta_1 = \frac{\left(\mathbf{c}_1^T\mathbf{h}_1\right)^2 P_1}{\mathbf{c}_1^T\mathbf{c}_1\sigma^2 + \sum_{k=2}^{K}\left(\mathbf{c}_k^T\mathbf{h}_k\right)^2 P_k}$$

is the minimum mean square error (MMSE) receiver. The SIR of user 1 is given by

$$\beta_1^{MMSE} = P_1\mathbf{h}_1^T\left(\mathbf{H}_1\mathbf{D}_1\mathbf{H}_1^T + \sigma^2\mathbf{I}\right)^{-1}\mathbf{h}_1$$

where

$$\mathbf{D}_1 = \text{diag}\left(P_2, \ldots, P_n\right) \in \mathbb{R}^{N\times(n-1)}, \quad \mathbf{H}_1 = \left(\mathbf{h}_2, \ldots, \mathbf{h}_n\right) \in \mathbb{R}^{N\times(n-1)}$$

Assume that the $\mathbf{h}_k$ are i.i.d. random vectors, each consisting of i.i.d. random variables with appropriate moments. Moreover, suppose that $N/n \to c > 0$ and $F_{\mathbf{D}_1}(x) \to H(x)$. Then, by Lemma 2.7 in [233] and the Helly–Bray lemma, we have

$$\beta_1^{MMSE} - P_1\int \frac{1}{x+\sigma^2} dF_{c,H}(x) \to 0, \qquad \text{in probability}$$

To compare the performance of different receivers we may then use the value of $\int \frac{1}{x+\sigma^2} dF_{c,H}(x)$ with the limiting SIR of the other linear receiver. However, we usually do not have an explicit expression for $F_{c,H}(x)$. Thus, we may use the kernel estimate $\int \frac{1}{x+\sigma^2} dF_n(x)$ to estimate $\int \frac{1}{x+\sigma^2} dF_{c,H}(x)$, with the help of (4.24). □

Example 4.3.7 (statistical inference of properties of the true covariance matrix) We use $f_n(x)$, defined in (4.16), to infer, in some way, some statistical properties of the true covariance matrix $\mathbf{T}_n$. Specifically speaking, by (4.11), we may evaluate the Stieltjes transform of the kernel estimator $f_n(x)$

$$m_{f_n}(z) = \int \frac{1}{x-z} f_n(x)\, dx, \qquad z \in \mathbb{C}^+ \tag{4.27}$$

By (4.13), we may then obtain $m_{g_n}(z)$ defined as

$$m_{g_n}(z) = -\frac{1-c}{z} + cm_{f_n}(z)$$

On the other hand, we conclude from (4.14) that

$$\frac{m_{g_n}(z)\left(c-1-zm_{g_n}(z)\right)}{c} = \int \frac{1}{t+1/m_{g_n}(z)} dH(t) \tag{4.28}$$

Note that $m_{g_n}(z)$ has a positive imaginary part. Therefore, with notation

$$z_1 = -1/m_{g_n}(z), \quad s\left(z_1\right) = \frac{m_{g_n}(z)\left(c-1-zm_{g_n}(z)\right)}{c}$$

we can rewrite (4.28) as

$$s\left(z_1\right) = \int \frac{1}{t-z_1} dH(t), \qquad z_1 \in \mathbb{C}^+ \tag{4.29}$$

As a result, in view of the inversion formula

$$F\left\{\left[a,b\right]\right\}=\frac{1}{\pi}\lim_{v\to\infty}\int_a^b \operatorname{Im} m_F\left(u+iv\right)du \tag{4.30}$$

we may recover $H(t)$ from $s(z_1)$ as given in (4.29). However, $s(z_1)$ can be estimated by the resulting kernel estimate $m_{f_n}(z)$ through the use of

$$\frac{m_{g_n}(z)\left(c-1-zm_{g_n}(z)\right)}{c},\quad \text{where } m_{g_n}(z)=-\frac{1-c}{z}+cm_{f_n}(z) \tag{4.31}$$

Once $H(t)$ is estimated, we may further estimate the functions of the true covariance matrix $\mathbf{T}_n$ such as $\frac{1}{n}\operatorname{Tr}\mathbf{T}_n^2$. Indeed, by the Helly–Bray lemma, we have

$$\frac{1}{n}\operatorname{Tr}\mathbf{T}_n^2=\int t^2dH_n(t)\xrightarrow{D}\int t^2dH(t)$$

where D stands for convergence in distribution. Thus, we may use an estimator for $\frac{1}{n}\operatorname{Tr}\mathbf{T}_n^2$, based on the resulting (4.31). We conjecture that the estimators of $H(t)$ and the corresponding functions like $\frac{1}{n}\operatorname{Tr}\mathbf{T}_n^2$ obtained by the above method are also consistent. A rigorous argument is currently being pursued by [228]. □

Example 4.3.8 (MATLAB simulations) We perform a simulation study to investigate the behavior of the kernel density estimators of the Marchenko–Pastur law. We consider one population with Gaussian distribution. The kernel is selected as

$$K(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$$

which is the standard normal density function. The bandwidth is chosen as $h=h(n)=n^{-\frac{2}{5}}$.

From each population, we generate two samples with sizes $N\times n$ equal to 50×200 and 800×3200, respectively. We can therefore form two random matrices, $\left(X_{ij}\right)_{50\times 200}$ and $\left(X_{ij}\right)_{800\times 3200}$. The kernel density estimator defined in (4.16) is

$$\hat{f}_n(x)=\frac{1}{N\times n^{-2/5}}\sum_{i=1}^{N}K\left(\left(x-\lambda_i\right)/n^{-2/5}\right)$$

where $\lambda_i, i=1,\ldots,N$, are eigenvalues of $\frac{1}{n}\left(X_{ij}\right)_{N\times n}\left(X_{ij}\right)_{N\times n}^{T}$. We consider two examples $n=200, N=50$, and 800×3200, respectively. This curve of $\hat{f}_n(x)$ is drawn and labeled as "Kernel Density Estimation" in Figure 4.2 and Figure 4.3, in comparison with the theoretical curve labeled as "Marchenko–Pastur."

Let us consider another case: the sum of several random matrices. Let $\mathbf{X}, Y$ be two independent random matrices of $N\times n$. Let Z be the random matrix of $N\times n$ whose entries are Bernoulli random variables. We consider the sample covariance matrix

$$\frac{1}{n}\left(\mathbf{X}+\mathbf{Y}+\sigma\mathbf{Z}\right)\left(\mathbf{X}+\mathbf{Y}+\sigma\mathbf{Z}\right)^T$$

where σ is the scaling parameter. The variance of the entries of $\mathbf{X}+\mathbf{Y}+\mathbf{Z}$ is normalized to 1. Figure 4.4 shows that the sum of independent random matrices will not affect the density function. The parameter σ also has no impact.

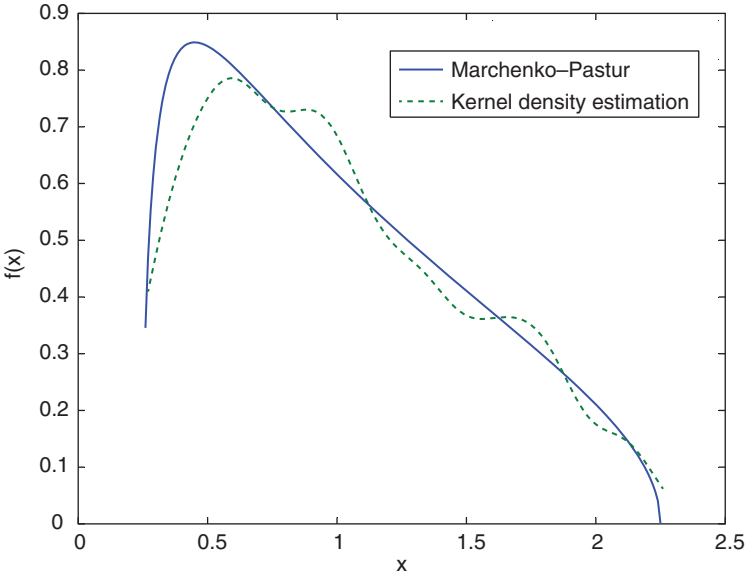


Figure 4.2 Spectral density curves for sample covariance matrices $\frac{1}{n}\left(X_{ij}\right)_{N\times n}\left(X_{ij}\right)_{N\times n}^{T}$, $N = 50$; $n = 200$. X_{ij} are i.i.d. standard Gaussian distribution with zero mean and variance 1, or $X_{ij} \sim \mathcal{N}(0, 1)$. In MATLAB: `X=randn(N,n);`

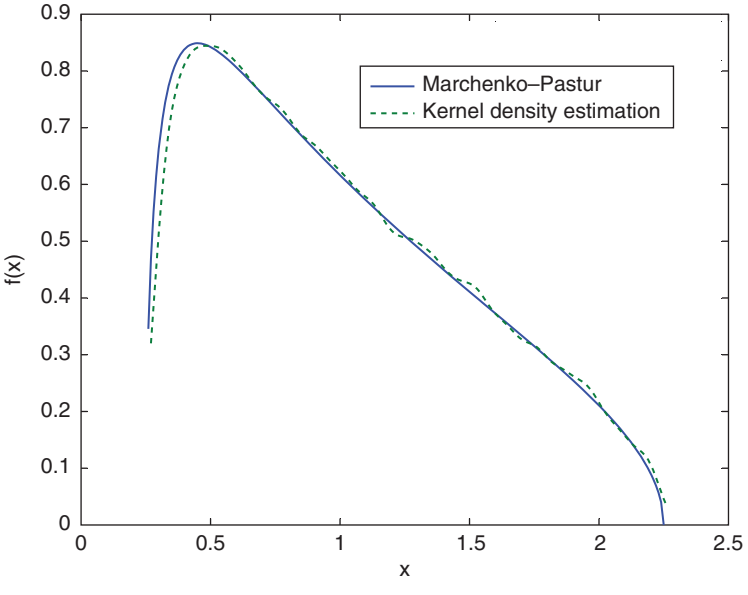


Figure 4.3 The same as Figure 4.2 except $N = 800$; $n = 3200$.

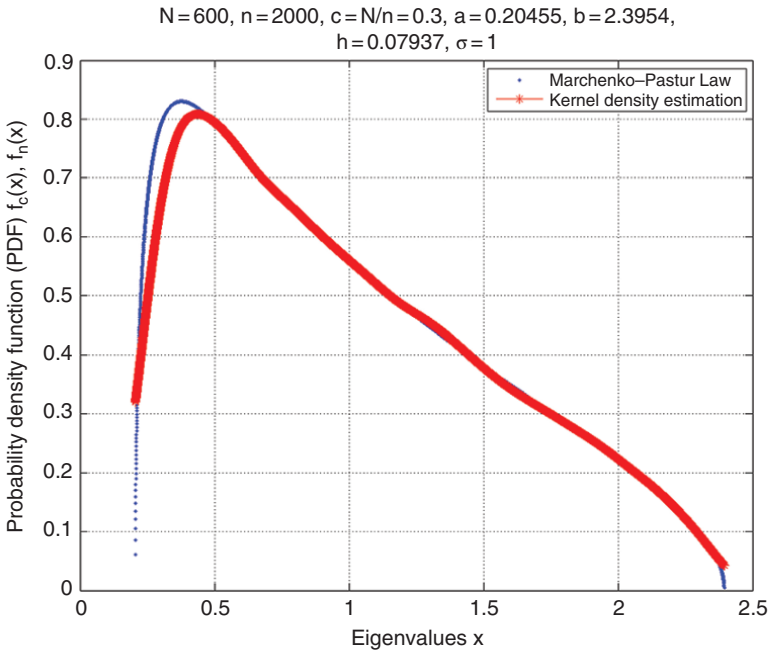


Figure 4.4 The kernel density estimation of $\frac{1}{n}(\mathbf{X}+\mathbf{Y}+\sigma\mathbf{Z})(\mathbf{X}+\mathbf{Y}+\sigma\mathbf{Z})^T$ is compared with the Marchenko-Pastur law. $N = 600$, $n = 2000$, $h = 1/n^{1/3}$, and $\sigma = 1$.

Although, sometimes, we do not know its exact formula, we can predict the limiting spectral density function. The kernel spectral density curve is consistent. The kernel spectral density estimator is robust with respect to the bandwidth selection.

The MATLAB code is included here for convenience.

```
clear all;
%Reference
% NONPARAMETRIC ESTIMATE OF SPECTRAL DENSITY FUNCTIONS OF
% SAMPLE COVARIANCE MATRICES: A FIRST STEP
% Bing-Yi Jing, Guangming Pan, Qi-Man Shao and Wang Zhou
% The Annals of Statistics, Vol. 38, No. 6, 3724-3750,2010.
N=50; n=200; h=1/n^(2/5);
c=N/n; a=(1-sqrt(c))^2; b=(1+sqrt(c))^2;
x=(a+0.01):0.01:b;
fcx=(1/2/pi/c./x).*sqrt((b-x).*(x-a));
      % the density function of Marcenko and Pastur law
X=randn(N,n); lambda=eig(1/n*X*X');
L=(b-a)/0.01; x1=a+0.01;
for j=1:L
for i=1:N
y=(x1-lambda(i))/h; Ky(i)=kernel(y);
end %N
```

```
fnx(j)=sum(Ky)/N/h; x1=x1+0.01; x2(j)=x1;
end %L
% figures
ifig=0;
ifig=ifig+1;figure(ifig)
plot(x,fcx,x2,fnx)
xlabel('x')
ylabel('f(x)')
legend('Marcenko-Pastur','Kernel Density Estimation');

function [Kx] = kernel(x)
Kx=1/sqrt(2*pi)*exp(-0.5*x.^2);
```

□

4.3.3 Central Limit Theorems for Kernel Estimators

The notation of this subsection is the same as in Section 4.3.2.

For a general $\mathbf{T}_n$, we refer to [229]. In this subsection, following [234], we consider the special case when $\mathbf{T}_n$ is the identity matrix, i.e. $\mathbf{T}_n = \mathbf{I}_n$. So we have

$$\mathbf{A}_n = \frac{1}{n}\mathbf{X}_n\mathbf{X}_n^T$$

It is equivalent to consider

$$\mathbf{B}_n = \frac{1}{n}\mathbf{X}_n^T\mathbf{X}_n$$

because the eigenvalues of $\mathbf{A}_n$ and $\mathbf{B}_n$ differ by $|n - N|$zero eigenvalues. The almost sure convergence of $F_{\mathbf{A}_n}(x)$ to the famous Marchenko–Pastur law (MP law) is fully understood under the 2nd moment condition of X_{11} when the dimension N is of the same order as the sample size n.

After establishing the strong law of large numbers, we may wish to prove the central limit theorem (CLT). However, even for the Wishart ensemble, there is no CLT available in the literature about $F_{\mathbf{A}_n}(\cdot)$ due to the shortage of powerful tools. Hence it is also impossible to make inference based on the individual eigenvalue of the sample covariance matrix when one only has finite moment conditions. These difficulties push us to seek other possible ways to make statistical inferences.

Using the notation

$$m(z) = m_{F_{\mathbf{A}_n}}(z)$$

which $m(z)$ satisfies (4.10), we obtain the relationship between Stieltjes transform of the limit of $F_{\mathbf{A}_n}(x)$ and $m_{F_{\mathbf{B}_n}}(z)$

$$m_{F_{\mathbf{B}_n}}(z) = -\frac{1-c}{z} + cm_{F_{\mathbf{A}_n}}(z) \tag{4.32}$$

which gives the equation satisfied by $m_{F_{\mathbf{B}_n}}(z)$

$$z = -\frac{1}{m_{F_{\mathbf{B}_n}}(z)} + \frac{c}{1 + m_{F_{\mathbf{B}_n}}(z)} \tag{4.33}$$

For the kernel function $K(\cdot)$ we assume that

$$\lim_{|x|\to\infty} |xK(x)| = \lim_{|x|\to\infty} |xK'(x)| = 0 \tag{4.34}$$

$$\int K(x)\,dx = 1, \quad \int |xK'(x)|\,dx < \infty, \quad \int |xK''(x)|dx < \infty \tag{4.35}$$

$$\int xK(x)\,dx = 0, \quad \int x^2 |K(x)| < \infty \tag{4.36}$$

Let $z = u + iv$, where $u \in \mathbb{R}$ and v is in a bounded interval, say $[?v_0, v_0]$ with $v_0 > 0$. Suppose that

$$\int_{-\infty}^{\infty} \left|K^{(j)}(z)\right| du < \infty, \quad j = 0, 1, 2 \tag{4.37}$$

uniformly in $v \in [?v_0, v_0]$, where $K^{(j)}(z)$ denotes the j-th derivative of $K(z)$. Also suppose that

$$\lim_{|x|\to\infty} \left|xK\left(x + iv_0\right)\right| = \lim_{|x|\to\infty} \left|xK'\left(x + iv_0\right)\right| = 0 \tag{4.38}$$

The distribution function the MP law is

$$\mathbb{F}_c(x) = \int_{-\infty}^{x} f_c(y) dy$$

where $f_c(x)$ is defined in (4.8). Our first result is the CLT for $\left(F_n(x) - \mathbb{F}_{c_n}(x)\right)$

Theorem 4.3.9 Suppose that

- $h = h(n)$ is a sequence of positive constants satisfying

$$\lim_{n\to\infty} \frac{nh^2}{\sqrt{\ln h^{-1}}} \to 0, \quad \lim_{n\to\infty} \frac{1}{nh^2} \to 0$$

- $K(x)$ satisfies (4.34)–(4.38) and is analytic on open interval including

$$\left[\frac{a-b}{h}, \frac{b-a}{h}\right]$$

- X_{ij} are i.i.d. with $\mathbb{E}X_{11} = 0$, $Var\left(X_{11}\right) = 1$, $\mathbb{E}X_{11}^4 = 3$ and $\mathbb{E}X_{11}^{32} < \infty$, $c_n \to c \in (0, 1)$

Then, as $n \to \infty$, for any fixed positive integer d and different points $x_1, \ldots, x_d$ in (a, b), the joint limiting distribution of

$$\frac{\sqrt{2}\pi n}{\sqrt{\ln n}} \left(F_n(x) - \mathbb{F}_{c_n}(x)\right) \sim \mathcal{N}\left(0, \mathbf{I}_d\right), \qquad j = 1, \ldots, d \tag{4.39}$$

is multivariate normal with mean zero and covariance matrix $\mathbf{I}_d$, the $d \times d$ identity matrix.

The convergence rate $n/\sqrt{\ln n}$ is consistent with the conjectured convergence rate $n/\sqrt{\ln n}$ of the ESD of sample covariance matrices to the Marchenko–Pastur law. It is easy to check that the Gaussian kernel $K(x) = (2\pi)^{-1/2} e^{-x^2/2}$ satisfies all conditions specified in Theorem 4.3.9.

Based on Theorem 4.3.9, we may further develop the smoothed quantile estimators of the Marchenko–Pastur law. For $0 \le \alpha < 1$, define the α-quantile of the Marchenko–Pastur law by

$$x_\alpha = \inf\left\{x, \mathbb{F}_{c_n}(x) > \alpha\right\} \tag{4.40}$$

and its estimator by

$$x_{n,\alpha} = \inf\left\{x, F_n(x) > \alpha\right\} \tag{4.41}$$

Under the assumptions of Theorem 4.3.9, we have

$$\frac{n}{\sqrt{\ln n}}\left(x_{n,\alpha} - x_\alpha\right) \to \mathcal{N}\left(0, \frac{1}{2\pi^2 f_c^2\left(x_\alpha\right)}\right), \quad x_\alpha \in (a,b) \tag{4.42}$$

where $f_c(x)$ and a, b are defined in (4.8).

The next theorem is the CLT for $f_n(x)$.

Theorem 4.3.10 Suppose that

- $h = h(n)$ is a sequence of positive constants satisfying

$$\lim_{n\to\infty} \frac{\ln h^{-1}}{nh^2} \to 0, \quad \lim_{n\to\infty} nh^3 = 0 \tag{4.43}$$

- $K(x)$ satisfies (4.34)–(4.38) and is analytic on open interval including

$$\left[\frac{a-b}{h}, \frac{b-a}{h}\right]$$

- X_{ij} are i.i.d. with $\mathbb{E}X_{11} = 0$, $Var\left(X_{11}\right) = 1$, $\mathbb{E}X_{11}^4 = 3$ and $\mathbb{E}X_{11}^{32} < \infty$, $c_n \to c \in (0,1)$.

Then, as $n \to \infty$, for any fixed positive integer d and different points $x_1, \ldots, x_d$ in (a,b), the joint limiting distribution of

$$nh\left(f_n\left(x_i\right) - f_{c_n}\left(x_i\right)\right) \sim \mathcal{N}\left(0, \sigma^2 \mathbf{I}_d\right), \quad i = 1, \ldots, d \tag{4.44}$$

is multivariate normal with mean zero and covariance matrix $\sigma^2\mathbf{I}_d$, where

$$\sigma^2 = -\frac{1}{2\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} K'\left(u_1\right)K'\left(u_2\right)\ln\left(u_1 - u_2\right)^2 du_1 du_2 \tag{4.45}$$

Here $f_c(x)$ and a, b are defined in (4.8). The Gaussian kernel $K(x) = (2\pi)^{-1/2}e^{-x^2/2}$ also satisfies all conditions specified in Theorem 4.3.10. Theorem 4.3.10 is actually a corollary of the following theorem.

Theorem 4.3.11 When the condition $\lim_{n\to\infty} nh^3 = 0$ in Theorem 4.3.10 is replaced by

$$\lim_{n\to\infty} h = 0$$

while the remaining conditions are unchanged, Theorem 4.3.10 holds as well if the random variables (4.44) are replaced by

$$nh\left(f_n\left(x_i\right) - \frac{1}{h}\int_a^b K\left(\frac{x_i - y}{h}\right) d\mathbb{F}_{c_n}(y)\right) \sim \mathcal{N}\left(0, \sigma^2\mathbf{I}_d\right), \quad x_i \in (a,b), \quad i = 1, \ldots, d$$

Example 4.3.12 (optimal bandwidth h) In practice, the bandwidth $h(n)$ as a function of n needs to be selected. First, we derive the theoretical result; then we show simulations.

We evaluate the quality of the estimate $f_n(x)$ by the mean integrated square error

$$\begin{aligned} L &= \mathbb{E}\left(\int_a^b \left(f_n(x) - f_{c_n}(x)\right)^2 dx\right) \\ &= \int_a^b \left(\text{Bias}\left(f_n(x)\right)\right)^2 dx + \int_a^b \text{Var}\left(f_n(x)\right) dx \end{aligned}$$

where $\text{Bias}\left(f_n(x)\right) = \mathbb{E}f_n(x) - f_{c_n}(x)$. It is easy to verify that (see [226] and [224])

$$\frac{1}{h}\int K\left(\frac{x-y}{h}\right) d\mathbb{F}_{c_n}(y) - f_{c_n}(x) = \frac{1}{2}h^2\left(f_{c_n}(x)\right)'' \int x^2 K(x)\,dx + O\left(h^3\right)$$

Although it is not rigorous from Theorem 4.3.11 we roughly have

$$\mathbb{E}f_n(x) - \frac{1}{h}\int K\left(\frac{x-y}{h}\right) d\mathbb{F}_{c_n}(y) = o\left(\frac{1}{nh}\right)$$

and

$$\text{Var}\left(f_n(x)\right) = \frac{\sigma^2}{n^2h^2} + o\left(\frac{\sigma^2}{n^2h^2}\right)$$

Recall that σ^2 is defined in (4.45). These give

$$L = \left[\frac{1}{2}h^2\left(f_{c_n}(x)\right)'' \int x^2 K(x)\,dx + O\left(h^3\right) + o\left(\frac{1}{nh}\right)\right]^2 + \frac{\sigma^2(b-a)}{n^2h^2} + o\left(\frac{\sigma^2}{n^2h^2}\right)$$

Differentiating the above with respect to h and setting it equal to zero, we see that the asymptotic optimal bandwidth is

$$h_\star = \left(\frac{\sigma^2(b-a)}{2n^2c_1^2}\right)^{1/6}$$

where $c_1 = \frac{1}{2}\left(f_{c_n}(x)\right)'' \int x^2 K(x)\,dx < \infty$. This is different from the asymptotic optimal bandwidth $O\left(1/n^{1/5}\right)$ in classical density estimates (see [226]).

Figure 4.5 illustrates (4.44) by showing how the kernel density estimation f_n deviates from the limiting Marchenko–Pastur law $f_{c_n}(x)$: $nh\left(f_n\left(x_i\right) - f_{c_n}\left(x_i\right)\right), i = 1, \ldots, d$. We have found that $h(n) = 1/n^{\frac{1}{3}}$ is the optimal bandwidth; this observation agrees with the theoretical derivation above. The case for $h(n) = 1/n^{\frac{1}{5}}$ is shown in Figure 4.6. ∎

4.3.4 Estimation of Noise Variance

Example 4.3.13 (MIMO channels) The observation vectors $\left\{\mathbf{x}_i\right\}_{1\leqslant i\leqslant n}$ are N-dimensional and satisfy the equation

$$\mathbf{x}_i = \mathbf{H}\mathbf{z}_i + \mathbf{w}_i + \boldsymbol{\mu}, \quad i = 1, \ldots, n \tag{4.46}$$

Here, $\mathbf{z}_i$ is an m-dimensional common factor where $m << N$, $\mathbf{H}$ is an $N \times m$ matrix, $\boldsymbol{\mu}$ represents the general mean, and $(\mathbf{w}_i)$ is a sequence of independent noise vectors. The random vectors $\mathbf{z}_i$ and the noise $\mathbf{w}_i$ have a Gaussian distribution and they are both unobserved. We have the choice

$$\mathbb{E}\mathbf{z}_i = 0 \text{ and } \mathbb{E}\mathbf{z}_i\mathbf{z}_i^T = \mathbf{I}; \quad \mathbf{R}_w = \text{cov}\left(\mathbf{w}_i\right) \text{ is diagonal}$$

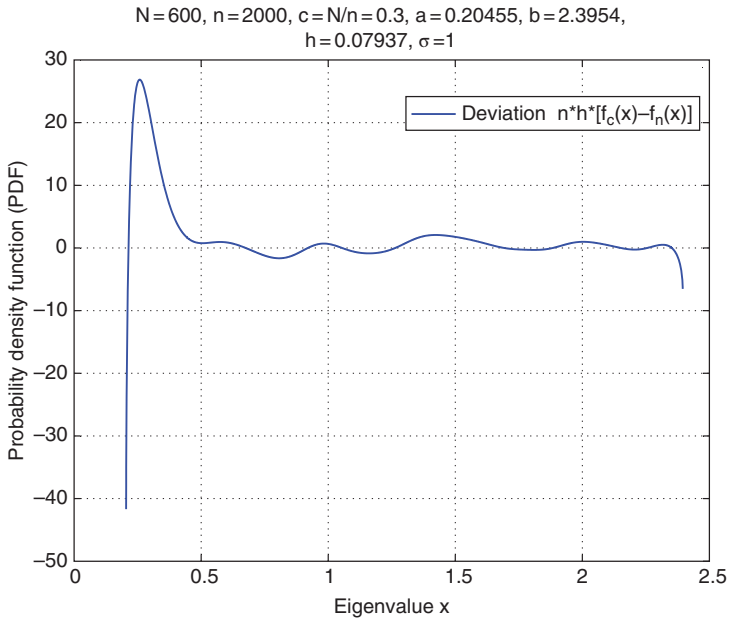


Figure 4.5 The kernel density estimation $f_n(x)$ deviates from the limiting Marchenko–Pastur law $f_{c_n}(x)$: $nh\left(f_n\left(x_i\right) - f_{c_n}\left(x_i\right)\right), i = 1, \ldots, d$. The optimal bandwidth $h = 1/n^{1/3}$. Here $d = 8763$ points are plotted. The setting is the same as Figure 4.2 unless otherwise specified.

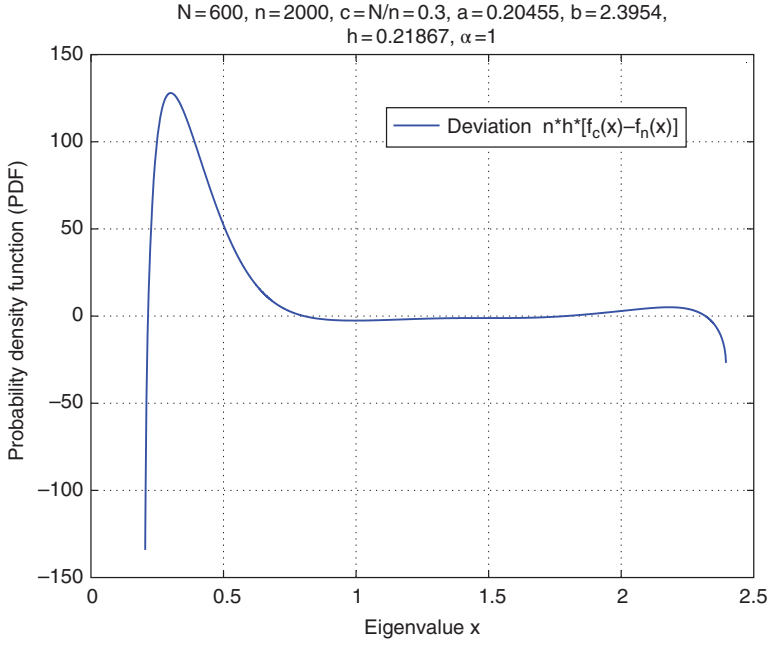


Figure 4.6 The same as Figure 4.5 except that $h = 1/n^{1/5}$.

and

$$\mathbf{\Gamma} := \mathbf{H}^T \mathbf{R}_w^{-1} \mathbf{H}$$

is diagonal with distinct diagonal elements. For a white Gaussian noise vector, we have $\mathbf{R}_w = \sigma^2 \mathbf{I}$. Therefore, the true covariance matrix of $\left\{\mathbf{x}_i\right\}_{1 \leqslant i \leqslant n}$ is

$$\mathbf{\Sigma} = \mathbf{H}\mathbf{H}^T + \mathbf{R}_w \tag{4.47}$$

The maximum likelihood estimator of $\boldsymbol{\mu}$ is $\overline{\mathbf{x}}$ and those of $\mathbf{\Gamma}$ and $\mathbf{R}_w$ are obtained by solving the following implicit equations

$$\mathbf{H}\left(\mathbf{\Gamma} + \mathbf{I}_m\right) = \mathbf{S}_n \mathbf{R}_w^{-1} \mathbf{H} \tag{4.48}$$

$$\operatorname{diag}\left(\mathbf{H}\mathbf{H}^T + \mathbf{R}_w\right) = \operatorname{diag}\left(\mathbf{S}_n\right), \quad \text{with } \mathbf{\Gamma} := \mathbf{H}^T \mathbf{R}_w^{-1} \mathbf{H} \text{ diagonal} \tag{4.49}$$

For the white Gaussian noise, the estimation of $\mathbf{R}_w = \sigma^2 \mathbf{I}$ is reduced to that of σ^2. (4.48) and (4.49) become

$$\mathbf{H}\left(\mathbf{\Gamma} + \mathbf{I}_m\right) = \frac{1}{\sigma^2} \mathbf{S}_n \mathbf{H}$$

$$N\sigma^2 = \operatorname{Tr}\left(\mathbf{S}_n - \mathbf{H}\mathbf{H}^T\right), \quad \text{with } \mathbf{\Gamma} := \frac{1}{\sigma^2} \mathbf{H}^T \mathbf{H} \text{ diagonal}$$

□

The maximum likelihood estimation is obtained as

$$\hat{\sigma}^2 = \frac{1}{N-m} \sum_{i=m+1}^{N} \lambda_i\left(\mathbf{S}_n\right) \tag{4.50}$$

and

$$\hat{\mathbf{H}}_i = \left(\lambda_{n,i} - \hat{\sigma}^2\right)^{1/2} \mathbf{v}_{n,i}, \quad 1 \leqslant i \leqslant m \tag{4.51}$$

where $\mathbf{v}_{n,i}$ is the normalized eigenvector of $\mathbf{S}_n$ corresponding to $\lambda_{n,i}$ $(1 \leqslant i \leqslant m)$, where λ_i is the eigenvalues of $\mathbf{S}_n$ the sample covariance matrix

$$\mathbf{S}_n = \frac{1}{n-1} \sum_{i=1}^{n} \left(\mathbf{x}_i - \overline{\mathbf{x}}\right)\left(\mathbf{x}_i - \overline{\mathbf{x}}\right)^T$$

and where $\overline{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$ the sample mean for the random vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^N$.

In the classical setting, hereafter called the low-dimensional setting, the asymptotic likelihood theory is developed by fixing the dimension N while the sample size $n \to \infty$. The maximum likelihood estimations are asymptotically normal with the standard $\sqrt{n}$-convergence rate. In particular, as $n \to \infty$

$$\sqrt{n}\left(\hat{\sigma}^2 - \sigma^2\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, s^2\right), \quad s^2 = \frac{2\sigma^4}{N-m} \tag{4.52}$$

Consider the so-called spiked true covariance matrix model

$$\operatorname{spec}(\mathbf{\Sigma}) = (\underbrace{\alpha_1, \ldots, \alpha_1}_{n_1}, \ldots, \underbrace{\alpha_K, \ldots, \alpha_K}_{n_K}, \underbrace{0, \ldots, 0}_{N-m}) + \sigma^2 \underbrace{(1, \ldots, 1)}_{N} \tag{4.53}$$

where (α_i) are non-null eigenvalues of $\mathbf{HH}^T$ with multiplicity numbers (n_i) satisfying $n_1 + \cdots + n_K = m$.

When the dimension N is large compared to the sample size n, the maximum likelihood estimation $\hat{\sigma}^2$ in (4.50) has a negative bias. Assuming some conditions are satisfied, we have

$$\frac{(N-m)}{\sigma^2\sqrt{2c}}\left(\hat{\sigma}^2-\sigma^2\right)+\beta\left(\sigma^2\right)\xrightarrow{\mathcal{L}}\mathcal{N}(0,1) \tag{4.54}$$

where $\beta\left(\sigma^2\right)=\sqrt{c/2}\left(m+\sigma^2\sum_{i=1}^{m}\frac{1}{\alpha_i}\right)$, and $c_n = p/(n-1)\to c>0$, as $n\to\infty$. Therefore, for high-dimensional data, the maximum likelihood estimation $\hat{\sigma}^2$ has an asymptotic bias $-\beta\left(\sigma^2\right)$ (after normalization). This bias is a complicated function of the noise variance and the m non-null eigenvalues of the matrix $\mathbf{HH}^T$. The above CLT (4.54) is still valid if $\tilde{c}_n = (N-m)/n$ is substituted for c. Now if we let $N << n$, so that $\tilde{c}_n \approx 0$ and $\beta\left(\sigma^2\right)\approx 0$, and thus

$$\frac{(N-m)}{\sigma^2\sqrt{2c}}\left(\hat{\sigma}^2-\sigma^2\right)+\beta\left(\sigma^2\right)\approx\frac{\sqrt{N-m}}{\sigma^2\sqrt{2}}\left(\hat{\sigma}^2-\sigma^2\right)$$

This is exactly the classical CLT (4.52) for known under the classical low-dimensional scheme. From this point of view, (4.54) gives a natural extension of the classical CLT (4.52) to the high-dimensional context.

4.4 Limiting Spectral Distribution of Time Series

Time series play a central role in the analysis of real-world applications.

4.4.1 Vector Autoregressive Moving Average (VARMA) Models

In this section we study the limiting spectral distribution of large-dimensional sample covariance matrices of a stationary and invertible VARMA (p,q) model. We study the limiting spectral distribution of a population covariance matrix and a sample covariance matrix for a VARMA (p,q) model. The relationship between the power spectral density function and limiting spectral distribution of large-dimensional covariance matrices of VARMA (p,q) is also established.

In the analysis of large-dimensional data, vector autoregressive moving average (VARMA) models are an important class of *linear* multivariate time-series models with a wide range of applications.

The power spectral density function of an ARMA(p,q) process is defined as:

$$\Phi(\omega)=\frac{\sigma^2}{2\pi}\frac{\left|\phi\left(e^{-j\omega}\right)\right|^2}{\left|\theta\left(e^{(-j\omega)}\right)\right|^2},\quad -\pi\leqslant\omega\leqslant\pi$$

where

$$\phi(t)=1-\phi_1 t-\cdots-\phi_p t^p,\quad \theta(t)=1+\theta_1 t+\cdots+\theta_q t^q$$

and σ^2 is constant variance.

4.4.2 General Linear Process

The results in Section 4.4.1 are limited to ARMA-type processes rather than the general linear process considered in this section.

Let $\{\mathbf{X}_i\}, i = 1, \ldots, n$ be a sequence of p-dimensional real-valued random vectors and consider the associated empirical covariance matrix

$$\mathbf{S}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^T \tag{4.55}$$

In this section, we consider another aspect of such Marchenko–Pastur type theorems by examining time-series observations instead of an i.i.d. sample. Let us first consider an univariate real-valued linear process

$$z_t = \sum_{k=0}^{\infty} \phi_k w_{t-k}, \quad t \in \mathbb{Z} \tag{4.56}$$

where (w_k) is real-valued and weakly stationary white noise with mean 0 and variance 1. The p-dimensional process $(\mathbf{X}_t)$ considered in this section will be made by p independent copies of the linear process (z_t), i.e. for $\mathbf{X}_t = \left(X_{1t}, \ldots, X_{pt}\right)^T$

$$X_{it} = \sum_{k=0}^{\infty} \phi_k w_{i,t-k}, \quad t \in \mathbb{Z} \tag{4.57}$$

where the p coordinate processes $\left\{\left(w_{1,t}, \ldots, w_{p,t}\right)\right\}$ are independent copies of the univariate error process $\{w_t\}$ in (4.56). Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be the observations of the time series at time epochs $t = 1, \ldots, n$. We are interested in the empirical spectral density of the sample covariance matrix $\mathbf{S}_n$ in (4.56).

We always employ an usual convention that for any complex number z, $\sqrt{z}$ denotes its square root with a non-negative imaginary part.

Theorem 4.4.1 Assume that the following conditions hold: (i) The dimensions $p \to \infty, n \to \infty$, and $p/n \to c \in (0, \infty)$. (ii) The error process has a fourth moment: $\mathbb{E}w_t^4 < \infty$. (iii) The linear filter $\left(\left(\phi_k\right)\right)$ is absolutely summable, i.e. $\sum_{k=0}^{\infty} \left|\phi_k\right| < \infty$. Then almost surely the empirical spectral density of $\mathbf{S}_n$ tends to a nonrandom probability distribution $F(x)$. Moreover, the Stieltjes transform $s = s(z)$ of $F(x)$ (as a mapping from $\mathbb{C}^+$ into $\mathbb{C}^+$) satisfies the equation

$$z = -\frac{1}{s(z)} + \frac{1}{2\pi}\int_0^{2\pi} \frac{1}{cs(z) + [2\pi\Phi(\omega)]^{-1}} d\omega \tag{4.58}$$

where $G(\omega)$ is the spectral density of the linear process (z_t):

$$\Phi(\omega) = \frac{1}{2\pi}\left|\sum_{k=0}^{\infty} \phi_k e^{j\omega k}\right|^2, \quad \omega \in [0, 2\pi) \tag{4.59}$$

We provide a numerical algorithm for the computation of the density function $h(x)$ of the LSD defined in (4.58) through its Stieltjes transform $s(z)$. We have

$$s(z) = \frac{1}{-z + g(s(z))}$$

with

$$g(s(z)) = \frac{1}{2\pi}\int_0^{2\pi} \frac{1}{cs(z) + (2\pi\Phi(\omega))^{-1}} d\omega$$

The algorithm below is of fixed-point type.

Algorithm

1) For a given real x, let ε be small enough positive value and set $z = x + i\varepsilon$.
2) Choose an initial value $s_0(z) = u + i\varepsilon$ and iterate for $k \geq 0$ the above mapping

$$s_{k+1}(z) = \frac{1}{-z + g\left(s_k(z)\right)}$$

until convergence and let $s_K(z)$ be the final value.
3) Define the estimate of the density function $h(x)$ to be

$$\hat{h}(x) = \frac{1}{\pi} \operatorname{Im} s_K(z)$$ ∎

It is well known that this iterated map has good contraction properties guaranteeing the convergence of the algorithm.

Example 4.4.2 (ARMA process) For simplicity, we consider the simplest causal ARMA(1,1) process for the coordinates:

$$z_t = \phi z_{t-1} + w_t + \theta z_{t-1}, \quad t \in \mathbb{Z}$$

where $|\phi| < 1$ and θ is real. The purpose is to find a simplified form of general equation (4.58). We have

$$\frac{1}{2\pi\Phi(\omega)} = \left|\frac{1 - \phi e^{j\omega}}{1 + \theta e^{j\omega}}\right|^2$$

and

$$I = \frac{1}{2\pi}\int_0^{2\pi} \frac{1}{cs(z) + (2\pi\Phi(\omega))^{-1}} d\omega = \frac{1}{2\pi i}\oint_{|\xi|=1} \frac{1}{cs(z) + \left|\frac{1-\phi\xi}{1+\theta\xi}\right|^2} \frac{d\xi}{\xi}$$

By a lengthy but elementary calculation of residues, we find that for an ARMA(1,1) process, the general equation (4.58) reduces to

$$z = -\frac{1}{s(z)} + \frac{\theta}{cs(z)\theta - \phi} - \frac{(\phi + \theta)(1 + \phi\theta)}{(cs(z)\theta - \phi)^2} \frac{\operatorname{sgn}(\operatorname{Im}(\alpha))}{\sqrt{\alpha^{2-4}}}$$

where

$$\alpha = \frac{cs(z)(1 + \theta^2) + 1 + \phi^2}{cs(z)\theta - \phi}$$

In order to compute the density function of the LSD $F(x)$, it is important to an explicit formula for the integral in (4.58) to implement numerical algorithms. □

4.4.3 Large Sample Covariance Matrices for Linear Processes

A typical object of interest in many fields is the sample covariance matrix $\frac{1}{n-1}\mathbf{X}\mathbf{X}^T$ of a data matrix $\mathbf{X} = \left(X_{i,t}\right), i = 1, \ldots, p; t = 1, \ldots, n$.

Our aim in this section is to obtain a Marchenko–Pastur-type result in the case where there is dependence within the rows of $\mathbf{X}$. More precisely, for $i = 1, \ldots, p$, the i-th row of $\mathbf{X}$ is given by a linear process of the form

$$\left(X_{i,t}\right)_{t=1,\ldots,n} = \left(\sum_{j=0}^{\infty} c_j Z_{i,t-j}\right)_{t=1,\ldots,n}, \quad c_j \in \mathbb{R}$$

Here, is an array of independent random variables that satisfies

$$\mathbb{E}Z_{i,t} = 0, \quad \mathbb{E}Z_{i,t}^2 = 1, \quad \nu_4 = \sup_{i,t} \mathbb{E}Z_{i,t}^4 < \infty \tag{4.60}$$

as well as the Lindeberg-type condition that, for each $\varepsilon > 0$

$$\frac{1}{pn}\sum_{i=1}^{p}\sum_{j=1}^{n} \mathbb{E}\left(Z_{i,t}^2 \mathrm{I}_{\left(Z_{i,t}^2 \geq \varepsilon n\right)}\right) \to 0, \quad \text{as } n \to \infty \tag{4.61}$$

Clearly, Eq. (4.61) is satisfied if all $\left\{Z_{i,t}\right\}$ are identically distributed.

The novelty of the result is that we allow for dependence within the rows, and that the equation for the Stieltjes transform $m_F(z)$ is given in terms of the spectral density

$$\Phi(\omega) = \sum_{k\in\mathbb{Z}} \phi_k e^{-j\omega k}, \quad \omega \in [0, 2\pi]$$

of the linear processes X_i only, which is the Fourier transform of the autocovariance function

$$\phi_k = \sum_{j=0}^{\infty} c_j c_{j+|k|}, \quad k \in \mathbb{Z}$$

4.4.4 Stationary Processes

We consider a stationary causal process $\left(X_k\right)_{k\in\mathbb{Z}}$ as follows: let $\left(W_k\right)_{k\in\mathbb{Z}}$ be a sequence of i.i.d. real valued random variables and let $g : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}$ be a measurable function such that, for any $k \in \mathbb{Z}$

$$X_k = g\left(\mathbf{w}_k\right), \qquad \mathbf{w}_k = \left(\cdots, W_{k-1}, W_k\right) \tag{4.62}$$

is a proper random variable, $\mathbb{E}g\left(\mathbf{w}_k\right) = 0$ and $\left\|g\left(\mathbf{w}_k\right)\right\|_2 < \infty$

The framework (4.62) is very general and it includes many widely used linear and nonlinear processes (see, for example, [235]). We refer to [236] for many examples of stationary processes that are of the form (4.62). Following [237] and [235]), $\left(X_k\right)_{k\in\mathbb{Z}}$ can be viewed as a physical system with $\mathbf{w}_k$ (respectively X_k) being the input (respectively output) and g being the transform or data-generating mechanism.

For a positive integer n, we consider n independent copies of the sequence $\left(W_k\right)_{k\in\mathbb{Z}}$ that we denote by $\left(W_k^{(i)}\right)_{k\in\mathbb{Z}}$, for $i = 1, \ldots, n$. Setting

$$\mathbf{w}_k^{(i)} = \left(\cdots, W_{k-1}^{(i)}, W_k^{(i)}\right), \quad X_k^{(i)} = g\left(\mathbf{w}_k^{(i)}\right)$$

it follows that

$$\left(X_k^{(1)}\right)_{k\in\mathbb{Z}}, \ldots, \left(X_k^{(n)}\right)_{k\in\mathbb{Z}}$$

are n independent copies of $\left(X_k\right)_{k\in\mathbb{Z}}$. Let $N = N(n)$ be a sequence of positive integers, and define for any $i \in \{1, \ldots, n\}$, $\mathbf{x}_i = \left(X_1^{(i)}, \ldots, X_N^{(i)}\right)$. Let

$$\mathbf{X}_n = \left(\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T\right) \in \mathbb{R}^{N\times n} \quad \text{and} \quad \mathbf{S}_n = \frac{1}{n}\mathbf{X}_n\mathbf{X}_n^T \in\in \mathbb{R}^{N\times N} \tag{4.63}$$

where $\mathbf{S}_n$ is the sample covariance matrix associated with $\left(X_k\right)_{k\in\mathbb{Z}}$. To derive the limiting spectral distribution of $\mathbf{S}_n$, we need to impose some dependence structure on $\left(X_k\right)_{k\in\mathbb{Z}}$. Let $\left(W_k^{'}\right)_{k\in\mathbb{Z}}$ be an independent copy of $\left(W_k\right)_{k\in\mathbb{Z}}$. We then define the functional dependence measure

$$\varepsilon(k) = \left\|X_k - X_k^{'}\right\|_2 \text{ for any integer } k \geqslant 0 \tag{4.64}$$

where $X_k^{'} = g\left(\mathbf{w}_k^{'}\right)$ with $\mathbf{w}_k^{'} = \left(\mathbf{w}_{-1}, W_0^{'}, W_1, \ldots, W_{k-1}, W_k\right)$. The coefficient $\varepsilon(k)$ measures how much the process will deviate, measured by the L^2 distance, from the original orbit $\left(g\left(\mathbf{w}_k\right)\right)_{k\geqslant 0}$ if we change the current input W_0 to an independent copy $W_0^{'}$. In addition, by Proposition 3 in [238], it satisfies

$$\left\|P_0\left(X_k\right)\right\|_2 \leqslant 2\varepsilon(k) \tag{4.65}$$

where for any k and j belonging to $\mathbb{Z}$ we have

$$P_j\left(X_k\right) = \mathbb{E}\left(X_k|\mathbf{w}_j\right) - \mathbb{E}\left(X_k|\mathbf{w}_{j-1}\right)$$

Theorem 4.4.3 Let $\left(X_k\right)_{k\in\mathbb{Z}}$ be defined in (4.62) and $\mathbf{S}_n$ (4.63). Assume that

$$\sum_{k\geqslant 0}\left\|P_0\left(X_k\right)\right\|_2 < \infty \quad \text{and} \quad \sum_{k\geqslant 0}\varepsilon^2(k) < \infty \tag{4.66}$$

and that $c(n) = N/n \to c \in (0, \infty)$. Then, with probability one, $F_{\mathbf{S}_n}(x)$ tends to a nonrandom probability distribution, whose Stieltjes transform $s(z)$ $(z \in \mathbb{C}^+)$ satisfies the equation

$$z = -\frac{1}{\underline{s}(z)} + \frac{c}{2\pi}\int_0^{2\pi}\frac{1}{\underline{s}(z) + \left(2\pi\Phi(\omega)\right)^{-1}}d\omega \tag{4.67}$$

where $\underline{s} = -\frac{1-c}{z} + cs(z)$ and $\Phi(\cdot)$ is the spectral density of $\left(X_k\right)_{k\in\mathbb{Z}}$.

Under the first part of condition (4.66), the series $\sum_{k\geqslant 0}\left|\mathrm{Cov}\left(X_0, X_k\right)\right|$ is finite. Thus (4.66) implies that the spectral density $\Phi(\cdot)$ of $\left(X_k\right)_{k\in\mathbb{Z}}$ exists, is continuous, and is bounded on $[0, 2\pi)$.

Condition (4.66) is referred in the literature as the Hannan–Heyde condition and is known to be sufficient for the validity of the central limit theorem for the partial sums (normalized by $\sqrt{n}$) associated with an adapted regular stationary process in L^2.

Consider functions of real-valued *linear* processes. Define

$$X_k = h\left(\sum_{i\geqslant 0} a_i W_{k-i}\right) - \mathbb{E}\left(h\left(\sum_{i\geqslant 0} a_i W_{k-i}\right)\right) \tag{4.68}$$

where $\left(a_i\right)_{i\in\mathbb{Z}}$ is a sequence of real numbers in ℓ^1 and is a sequence of i.i.d. random variables in L^1. We can give sufficient conditions in terms of the regularity of the function $h(x)$ for the condition (4.66) to be satisfied. See [239] for details.

4.4.5 Symmetrized Auto-cross Covariance Matrix

Consider the limiting spectral distribution (LSD) of a symmetrized autocross covariance matrix

$$\mathbf{R}_\tau = \frac{1}{2T}\sum_{k=1}^{T}\left(\mathbf{w}_k\mathbf{w}_{k+\tau}^H + \mathbf{w}_k^H\mathbf{w}_{k+\tau}\right)$$

where $\mathbf{w}_k = \left(W_{1k},\ldots,W_{Nk}\right)^T$ and $\left\{W_{it}\right\}$ are independent random variables with mean 0 and variance σ^2. Here, $\tau \geq 1$ denotes the number of lags. The motivation for this section comes from any large-dimensional model with a lagged time series structure that is central to large dimensional dynamic factor models [240] and singular spectrum analysis [241, 242].

Consider the framework of a large-dimensional dynamic k-factor model with lag q to understand the underlying motivation of this work. This takes the following form

$$\mathbf{y}_t = \sum_{i=0}^{q}\mathbf{H}_i\mathbf{x}_{t-i} + \mathbf{w}_t, \quad t = 1,\ldots,T$$

where $\mathbf{H}_i$s are $N\times k$ nonrandom matrices with full rank. For $t = 1,\ldots,T$, $\mathbf{x}_t$s are k-dimensional vectors of i.i.d. standard complex components with finite fourth moment and $\mathbf{w}_t$ are N-dimensional vectors of i.i.d. standard complex components with finite second moment, independent of $\mathbf{w}_t$. This model can be viewed as a large-dimensional information-plus-noise-type model [243], with information contained in the summation part and noise in $\mathbf{w}_t$. Here, "large dimension" refers to N and T, while the number of factors k and the number of lags q are small and fixed. Under this high-dimensional setting, an important statistical problem is the estimation of k and q.

For $\tau = 0$, $\mathrm{Cov}\left(\mathbf{x}_t\right) = \mathbf{\Sigma}_x$, the population covariance matrix of $\mathbf{y}_t$ has the same eigenvalues as those of

$$\begin{pmatrix} \sigma^2\mathbf{I} + \mathbf{H}^H\mathbf{\Sigma}_x\mathbf{H} & 0 \\ 0 & \sigma^2\mathbf{I} \end{pmatrix}$$

with the two diagonal blocks of size $k(q+1)\times k(q+1)$ and $(N-k(q+1))\times(N-k(q+1))$, respectively. Therefore, we have the spiked population model framework.

The limiting spectral distribution of $\mathbf{R}_\tau$ denoted as $F_\tau(x)$ has been derived: $F_\tau(x) = \lim_{N\to\infty} F_{\mathbf{R}_\tau}(x)$. See [244] for details.

Let us summarize the work of [245] in this context. The focus is on a class of time series known as linear processes (or MA(∞) processes) given by the representation

$$\mathbf{x}_t = \sum_{\ell=0}^{\infty}\mathbf{A}_\ell\mathbf{z}_{t-\ell}, \quad t\in\mathbb{Z}, \tag{4.69}$$

where $\left(\mathbf{A}_\ell : \ell \in \mathbb{N}\right)$ are $p \times p$ matrices, $\mathbf{A}_0$ the identity matrix, and $\left(\mathbf{z}_t : t \in \mathbb{Z}\right)$ are p-dimensional random vectors (innovations) with i.i.d. entries Z_{tj} (real or complex valued) with zero mean, unit variance and finite fourth moment. It is assumed that the matrices $\mathbf{A}_\ell$ are symmetric (in the real case) or Hermitian (in the complex case) and are simultaneously diagonalizable. Moreover, the stability requirement $\sum_{\ell=1}^{\infty} \ell \left\|\mathbf{A}_\ell\right\| < \infty$ is imposed, where $||\cdot||$ denotes the operator norm. These assumptions imply that, up to an unknown rotation, the coordinates of the process $\mathbf{x}_t$ are uncorrelated stationary linear processes with short range dependence. The goal is to relate the behavior of the ESD of the lag-τ symmetrized sample autocovariances, defined as

$$\mathbf{C}_\tau = \frac{1}{2n} \sum_{t=1}^{n-\tau} \left(\mathbf{x}_t \mathbf{x}_{t+\tau}^H + \mathbf{x}_{t+\tau} \mathbf{x}_t^H\right),$$

to that of the spectra of the coefficient matrices $\left(\mathbf{A}_\ell : \ell \in \mathbb{N}\right)$ when $p, n \to \infty$ such that $p/n \to c \in (0, \infty)$.

The class of models under study here includes the class of causal autoregressive moving average (ARMA) processes of finite orders satisfying the requirement that the coefficient matrices are simultaneously diagonalizable and the joint empirical distribution of their eigenvalues (when diagonalized in the common orthogonal or unitary basis) converges to a finite dimensional distribution. The results are expressed in terms of the Stieltjes transform of the ESD of the sample autocovariances.

4.4.6 Large Sample Covariance Matrices with Heavy Tails

This section deals with symmetric random matrices whose upper diagonal entries are obtained from a linear random field with heavy tailed noise. Our goal here is to weaken the moment conditions by allowing for heavy tails, and the assumption of independent entries by allowing for dependence within the rows and columns. Potential applications arise in portfolio management in finance, massive MIMO, and smart grid, where observations typically have heavy tails and dependence.

In the statistical analysis of high-dimensional data, we often try to reduce its dimensionality, while preserving as much of the variation in the data as possible. One important example of such an approach is the principal component analysis (PCA). The variances of the first k principal components are given by the k-largest eigenvalues of the covariance matrix. In practice, the true underlying covariance matrix is not available, thus one usually replaces it with the sample covariance matrix $\frac{1}{n}\mathbf{X}\mathbf{X}^T$, where $\mathbf{X}$ is a $p \times n$ data matrix. To account for large high-dimensional data sets, we study the k-largest eigenvalues of the sample covariance matrix when both the dimension of the data as well as the sample size go to infinity.

There are two cases: (i) observations are regularly varying with tail index $\alpha \in (0, 2)$—observations with infinite variance; (ii) observations have finite variances but infinite fourth moments with tail index $\alpha \in [2, 4)$. In many applications where data typically exhibits heavy tails, like finance, the assumption of an infinite variance might be too strong. So here we assume case (ii). This assumption is also consistent with the motivation to derive a theoretical framework for the use of PCA for high-dimensional data.

We assume that $\mathbf{X}$ is a $p \times n$ matrix with entries

$$X_{it} = \sum_{j=-\infty}^{\infty} c_j Z_{i,t-j}, \quad j \in \mathbb{N} \tag{4.70}$$

where the sequence c_j is absolutely summable, $\sum_{j=-\infty}^{\infty} \left|c_j\right| < \infty$, and $\left(Z_{it}\right)_{i,t}$ is an array of i.i.d. mean zero random variables with marginal distribution that is regularly varying with tail index $\alpha \in [2, 4)$, and *normalizing sequence* a_n:

$$\mathbb{E}Z_{11} = 0 \quad \text{and} \quad \lim_{n\to\infty} n\mathbb{P}\left(\left|Z_{it}\right| > a_n x\right) = x^{-\alpha}, \quad \text{for each } x > 0 \tag{4.71}$$

In other words, for each $i \in \mathbb{N}$, $\left(X_{it}\right)_t$ there is an infinite order moving average process driven by some regularly varying noise with finite variance but infinite fourth moment. From classical extreme value theory, the sequence a_n is necessarily characterized by

$$a_n = n^{1/\alpha} L(n) \tag{4.72}$$

for some slowly varying function $L : \mathbb{R}^+ \to \mathbb{R}^+$: a function with the property that, for each $x > 0$, $\lim_{t\to\infty} L(tx)/L(t) = 1$. Moreover we assume that Z_{11} satisfies the tail balancing condition given by

$$\lim_{x\to\infty} \frac{\mathbb{P}\left(Z_{11} > x\right)}{\mathbb{P}\left(\left|Z_{11}\right| > x\right)} = q = 1 - \lim_{x\to\infty} \frac{\mathbb{P}\left(Z_{11} \leqslant -x\right)}{\mathbb{P}\left(\left|Z_{11}\right| > x\right)} \tag{4.73}$$

for some $0 \leq q \leq 1$.

Definition 4.4.4 The (normalized) sample covariance matrix of the sample $\mathbf{X}$ is defined as the $p \times p$ matrix

$$\mathbf{S}_n = \frac{1}{a_{np}^2}\left(\mathbf{X}\mathbf{X}^T - n\mu_X \mathbf{I}_p\right)$$

where $\mu_X = \mathbb{E}Z_{11}^2 \sum_j c_j^2$ and $\mathbf{I}_p$ is the $p \times p$ identity matrix. We denote by $\lambda_1, \ldots, \lambda_p$ the unordered, and $\lambda_{(1)}, \ldots, \lambda_{(p)}$ by the ordered eigenvalues of $\mathbf{S}$.

It can be shown that $\mathbf{X}\mathbf{X}^T$ is dominated by its diagonal entries. If $\alpha > 2$, the diagonal entries have a finite mean $n\mu_X = n\mathbb{E}Z_{11}^2 \sum_j c_j^2$, which has to be subtracted in order to obtain a nontrivial limiting result.

If $\alpha = 2$, it is possible that $\mathbb{E}Z_{11}^2 = \infty$. In this case we replace μ_X in the above definition with the sequence of truncated means $ux^n = \sum_j c_j^2 \mathbb{E}\left(Z_{11}^2 \mathbf{I}_{\left\{Z_{11}^2 \leqslant a_{np}^2\right\}}\right)$.

The following theorem is a generalization of [246] to nonindependent entries, except that [246] assumes that p/n goes to some positive finite constant, while we assume that p is bounded by some small power of n. The random probability measure of its eigenvalues is defined as $\frac{1}{p}\sum_{i=1}^{p} \delta_{n^{-1}\lambda_i}$, where δ denotes the Dirac measure.

Theorem 4.4.5 Define the matrix $\mathbf{X} = \left(X_{it}\right)$ as in equations (4.70), (4.71)and (4.73) with $\alpha \in [2, 2)$. Suppose $n \to \infty, p\infty$ such that

$$\limsup_{n\to\infty} \frac{p_n}{n^\beta} < \infty \tag{4.74}$$

for some $\beta > 0$ satisfying

- $\beta < \max\left\{\frac{1}{3}, \frac{4-\alpha}{4(\alpha-1)}\right\}$ if $2 \leqslant \alpha < 3$, or
- $\beta < \frac{4-\alpha}{3\alpha-4}$ if $3 \leqslant \alpha < 4$.

Then the point process $N_n = \sum_{i=1}^{p} \delta_{\lambda_i}$ of the eigenvalues of $\mathbf{S}_n$ converges in distribution to a Poisson point process N_n with intensity measure ν, which is given by

$$\nu\left((x, \infty]\right) = \mathbb{E}N_n\left(x, \infty\right] = x^{-\alpha/2}\left|\sum_j c_j^2\right|^{\alpha/2}, x > 0$$

In particular, the theorem shows that the k largest eigenvalues $\lambda_{(1)} \geqslant \cdots \geqslant \lambda_{(k)}$ of $\mathbf{S}_n$, the variances of the k-largest principal components, converge jointly to a random vector with a distribution that only depends on k, the tail index α and the coefficients (c_j). Let (Y_i) be an i.i.d. sequence of exponentially distributed random variables with mean 1, $\mathbb{P}\left(Y_i > x\right) = e^{-x}$ for $x > 0$, and denote by $\Gamma_i = Y_1 + \cdots + Y_i$ their successive sum. Then we have that

$$\left(\lambda_{(1)}, \ldots, \lambda_{(p)}\right) \xrightarrow[n\to\infty]{D} \left(\Gamma_1^{-2/\alpha}, \ldots, \Gamma_k^{-2/\alpha}\right)\left(\sum_j c_j^2\right)$$

Bibliographical Remarks

In Section 4.1, we have drawn material from [247] and [245].

We followed [201] for the exposition of generalized Marchenko–Pastur distributions in Section 4.2.

We followed [248] in Section 4.3.4.

In Section 4.3.3, we followed [234] and [229]. Reference [249] is relevant. Sample covariance matrices (with/without empirical centering) are denoted by $\mathbf{S}$ and $\mathcal{S}$. It is proved in [249] that central limit theorems of eigenvalue statistics of $\mathbf{S}$ and $\mathcal{S}$ are different as $n \to \infty$ with N/n approaching a positive constant. Moreover, it is also proved that such a different behavior is not observed in the average behavior of eigenvectors [229].

Section 4.4.1 is taken from [250].

Section 4.4.2 is taken from [251].

Section 4.4.4 is taken from [239].

Section 4.4.5 is taken from [244].

Section 4.4.3 is taken from [252, 253].

Section 4.4.6 is taken from [253] with some minor notation changes to fit our habits. See also [146] and [254].

The Marchenko–Pastur law is treated for a linear time series in [245].

5

Large Hermitian Random Matrices and Free Random Variables

Finding *correlations* between observables is at the heart of scientific methodology. Once correlations between "causes" and "effects" are empirically established, we can start devising theoretical models to understand the mechanisms underlying such correlations, and use these models for prediction purposes. In many cases the number of possible causes and of resulting effects are both large. For example, in an industrial setting, one can monitor a large number of characteristics of a device (engine, hardware, etc.) during the production phase and correlate these with the performances of the final product. In economics and finance, one aims to understand the relation between a large number of possibly relevant factors. Nowadays, the number of macroeconomic time series available to economists is huge. This has led Granger [255] and others to suggest that "**large models**" should be at the forefront of the econometrics agenda. The idea of exploiting "large models" is the unified theme of this whole book. As a result, large random matrices are natural building blocks in the theoretical framework. As large random matrices can be regarded as free random variables, matrix-valued free probability theory is therefore relevant.

Random matrices find ubiquitous applications in many branches of science. The reason for this is twofold. First, random matrices possess a great degree of universality: eigenvalue properties of large matrices do not depend on details of the underlying statistical matrix ensemble. Second, random matrices can be viewed as *non commuting random variables*. As such, they form a basis of a non commutative probability theory where the whole matrix is treated as an element of the probabilistic space. In the limit when the size of the matrix tends to infinity, $N \to \infty$, the connection to the probability theory is becoming exact in the mathematical sense. This is the celebrated free-probability theory, where independent matrices play the role of free random variables (hereafter FRV). The most fundamental observation is that *data sets are usually organized as large matrices*. Large random matrices are regarded as free random variables, which is the basis for this chapter.

Our basic task is to represent these large data sets—data modeling for big data. Our basic approach is to learn from physicists. Physicists concentrated on the analysis of experimental data using tools borrowed from the analysis of real-world complex systems, to become a predictive theory at a high confidence level.

Often the first dimension of these matrices is equal to the number of degrees of freedom, N, and the second to the number of measurements, T. Typical examples are large economic/financial systems, sensing networks, wireless networks, and complex biological systems.

Smart Grid using Big Data Analytics: A Random Matrix Theory Approach, First Edition.
Robert C. Qiu and Paul Antonik.

A common feature of these methods is that they require massive accumulation and analysis of data, which is usually contaminated by statistical noise. Due to the high dimensionality of the system in question, its complex nature, nonlinearity, potential non stationarity and emerging collective behavior, the problem at hand becomes hard to solve using the traditional methods of multivariate statistical analysis. The basic methodology is to borrow ideas from emergent domains of physics and mathematics such as statistical theory of networks, percolation theory, spin glasses, random matrix theory, free random probability, and game theories.

In the language of a matrix-valued probability calculus, the quantum nature comes from the fact that basic objects of the probability calculus are operators, written as large, non commuting matrices, represented in economy by arrays of big data—in financial systems, wireless networks, sensor networks, smart grid, and so forth. The relevant observables in this language are related to the statistical properties of their spectra.

5.1 Large Economic/Financial Systems

Our basic approach to big data is the method of analogy, with the goal of parallel applications in financial systems, wireless networks, sensor networks, smart grid, and so forth. The concepts of statistical physics can enrich this science of big data, hopefully even making a major impact at the fundamental level.

Two conceptual revolutions were caused by Boltzmann (concepts of probability) and quantum mechanics (matrix-valued probability).

The concept of a random walk was formulated using the assumption of the Gaussian character of a stochastic process. Today, for a physicist, familiar with critical phenomena, the concept of a power law and large fluctuations is rather obvious. The second major factor changing the Gaussian world was the computer. The performance of computers has increased by several orders of magnitude. This fact has had a large impact on the economy. First, the speed and the range of transactions has changed drastically. In such a way, a computer started involuntarily to serve as an amplifier of fluctuations. Second, the economies and markets started to watch each other more closely, because computers allowed for *collecting exponentially more data*.

Since the 1990s, there has been a tendency for physicists to study the economy scientifically. One benefit of using computers was that economic systems started to save more and more data. Today markets collect an incredible amount of data (they remember practically every transaction). This triggers the need for new methodologies able to manage the data—data management for big data. In particular, the data started to be modeled and analyzed using methods—big data modeling and learning—borrowed widely from physics. These studies were devoted mostly to quantitative finance. To a large extent, they were triggered by the *vast amount of data accessible in this field*. In the science of big data, this is the reason why the state of the art for financial systems is ahead of other branches such as wireless networks, sensor networks and smart grid. The study of financial systems is more difficult than these big physical systems such as wireless networks, sensor networks, and smart grid, in that the latter can be controlled for data collection.

We must deal with large-scale phenomena in economic systems. There are too many random factors for us to consider. The ultimate goal is to reach a level of

understanding that would permit us to predict the reaction of the system to the change of macro economic parameters in the future.

It is worthwhile stressing how natural and fundamental is the use of large random matrices from the point of view of non commutative probability and central limit theorems. Thus, it is puzzling how late large random matrices (in our language matrix probabilities) were used for the analysis of financial data. The breakthrough came in 1999, as treated in Section 2.10.2. Matrix probability theory seems to be ideally suited for a better understanding the role of covariance matrices and a way to quantitatively assess the role of the noise, relevant correlations and the stability of the analysis. In our opinion, the full power of random matrix techniques has not yet been recognized by the quantitative finance community, not to mention by the big data community.

5.2 Matrix-Valued Probability

The basic building block for matrix-valued probability is the complex random matrix $\mathbf{X}$ of $N \times T$, $\mathbf{X} \in \mathbb{C}^{N \times T}$. Can one formulate an analog of the central limit theorem, if random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$ forming the sums

$$\mathbf{S}_n = \mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_n \tag{5.1}$$

do not commute? In other words, we are looking for a theory of probability that is noncommutative—$\mathbf{X}_i$s can be viewed as operators—but which should exhibit close similarities to the "classical" theory of probability.

Abstract operators may have matrix representations. If such constructions exist, we would have a natural tool for formulating probabilistic analysis directly in the space of matrices. Contemporary financial markets are characterized by collecting and processing enormous amount of data. Statistically, they may obey the matrix-valued central limit theorems. Matrix-valued probability theory is then ideally suited for analyzing the properties of large arrays of data. It also allows the reformulation of standard multivariate statistical analysis of covariance into a novel and powerful language. Spectral properties of large arrays of data may also provide a rather unique tool for studying chaotic properties, unraveling correlations and identifying unexpected patterns in very large sets of data.

The origins of noncommutative probability is linked with abstract studies of von Neumann algebras carried out in the 1980s. A new twist was given to the theory when it was realized that noncommuting abstract operators, called free random variables, can be represented as infinite matrices [126]. The concept of free random variables only very recently started to appear explicitly in physics [127, 128, 131].

In this section, we abandon a formal way and we shall follow the intuitive approach.

Our main goal is to study the *spectral* properties of large arrays of data, with big physical systems in mind. As large stochastic matrices obey central limit theorems with respect to their *measure*, spectral analysis is a powerful tool for establishing a stochastic feature of the whole set of matrix-ordered data simply by comparing their spectra to the analytically known results of random matrix theory. The deviations of empirical spectral characteristics from the spectral correlations of purely stochastic matrices can be used simultaneously as a source for inferring the relevant correlations, which are not so visible when investigated by other methods.

Let us assume, that we want to study the statistical properties of infinite random matrices. We are interested in the spectral properties of an $N \times N$ matrix $\mathbf{X}$ (in the limit $N \to \infty$), which is drawn from a matrix measure

$$\exp\left(-N \operatorname{Tr} V(\mathbf{X})\right) d\mathbf{X} \tag{5.2}$$

with a potential $V(\mathbf{X})$ (in general not necessarily polynomial). We shall restrict ourselves to complex Hermitian matrices for the moment because their spectrum is real. The average spectral density of the matrix $\mathbf{X}$ is defined as

$$\rho(\lambda) = \frac{1}{N} \langle \operatorname{Tr} \delta(\lambda - \mathbf{X}) \rangle = \frac{1}{N} \left\langle \sum_{i=1}^{N} \delta\left(\lambda - \lambda_i\right) \right\rangle \tag{5.3}$$

where $\langle \cdots \rangle$ means averaging over the ensemble (2.32). Then, $G(z)$ is a meromorphic function whose poles are on the real axis and correspond to the eigenvalues of the particular $\mathbf{X}$ matrix one is considering. Conversely, when the averaging operation is actually performed, and the $N \to \infty$ limit is taken, the poles of the Green's function start to merge into continuous intervals of the real line. In this limit, the Green's function becomes a holomorphic function everywhere in the complex plane except for the intervals mentioned above. Remarkably, those are the intervals where the eigenvalue density (2.33) is actually defined.

Using the standard folklore, that the spectral properties are related to the discontinuities of the Green's function, we may introduce

$$G(z) = \frac{1}{N} \left\langle \operatorname{Tr} \left(z\mathbf{I}_N - \mathbf{X}\right)^{-1} \right\rangle \tag{5.4}$$

where z is a complex variable and $\mathbf{I}_N$ is the identity matrix of $N \times N$. Due to the known properties of the distributions (Sokhotsky's formula)

$$-\frac{1}{\pi} \lim_{\varepsilon \to 0+} \operatorname{Im} G(z)\big|_{z=\lambda+i\varepsilon} = \rho(\lambda) \tag{5.5}$$

where PV stands for the principal value, we see that the imaginary part of the Green's function reconstructs spectral density (2.33)

$$-\frac{1}{\pi} \lim_{\varepsilon \to 0+} G(z)\big|_{z=\lambda+i\varepsilon} = \rho(\lambda) \tag{5.6}$$

so the Green's function, in the infinite matrix limit, is equivalent to the eigenvalue density and encodes all of the spectral density of the matrix ensemble under study. The whole framework is motivated by the fundamental relation of (2.36). Being holomorphic everywhere in the complex plane except for some cuts on the real line, the Green's function can typically be expanded into a *power series* around infinite z whose coefficients can be shown to be given by the following expression:

$$G(z) = \sum_{k=0}^{\infty} \frac{M_k}{z^{k+1}}, \quad M_k = \int \rho(\lambda) \lambda^k d\lambda \tag{5.7}$$

The m_k are called matrix moments and are usually summed up in the M-transform, or moment generating function

$$M_{\mathbf{X}}(z) = zG_{\mathbf{X}}(z) - 1 = \sum_{k=1}^{\infty} \frac{M_k}{z^{k+1}} \tag{5.8}$$

The natural Green's function will serve as an auxiliary construction that explains the crucial concepts of the theory of matrix-valued (noncommutative) probability theory. Let us define a functional inverse of the Green's function (sometimes called a Blue's function [128]):

$$G[B(z)] = z$$

The fundamental object in noncommutative probability theory, the so-called R function or R-transform, is defined as

$$R(z) = B(z) - \frac{1}{z} \tag{5.9}$$

With the help of the R-transform, we shall now uncover several astonishing analogies between classical and matrix-valued probability theory.

We shall start from the analog of the central limit theorem [126]: the spectral distribution for the sum of independent variables $\mathbf{X}_i, i = 1, \ldots, K$

$$\mathbf{S}_n = \frac{1}{\sqrt{n}} \left(\mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_n \right) \tag{5.10}$$

each with arbitrary probability measure with zero mean and finite variance $\left\langle \mathrm{Tr}\, \mathbf{X}_i \mathbf{X}_i^H \right\rangle = \sigma^2, i = 1, \ldots, n$, converges towards the distribution with R-transform $R(z) = \sigma^2 z$.

Let us now find the exact form of this limiting distribution. $R(z) = \sigma^2 z$, so it follows from (2.37) that $B(z) = \sigma^2 z + 1/z$, and its functional inverse fulfills

$$z = \sigma^2 G(z) + \frac{1}{G(z)} \tag{5.11}$$

The solution of this quadratic equation (with proper asymptotics $G(z) \to 1/z$ for large z) is

$$G(z) = \frac{z - \sqrt{z^2 - 4\sigma^2}}{2\sigma^2} \tag{5.12}$$

so the spectral density, supported by the cut of the square root, is

$$\rho(\lambda) = \frac{1}{2\pi\sigma^2} \sqrt{4\sigma^2 - \lambda^2} \tag{5.13}$$

This is the famous Wigner semicircle [102] (actually, semiellipse) ensemble. The omnipresence of this ensemble in various physical applications finds a natural explanation—it is a consequence of the central limit theorem for noncommuting random variables. Thus, the Wigner ensemble is a noncommutative analog of the Gaussian distribution in classical commutative probability. Indeed, one can show, that the measure (5.2) corresponding to the Green's function (5.12) is $V(\mathbf{X}) = \frac{1}{\sigma^2}\mathbf{X}^2$ for the real matrix case and $V(\mathbf{X}) = \frac{1}{\sigma^2}\mathbf{X}\mathbf{X}^H$ for the complex matrix case.

5.2.1 Eigenvalue Spectra for the Covariance Matrix and its Estimator

Statistical systems with many degrees of freedom appear in numerous research areas, such as big physical data systems. One of the most fundamental issues is the determination of correlations. In practice, we sample the system many times by carrying out independent measurements. For each sample, we estimate values of the elements of the covariance matrix, and then take the average over a set of samples. The statistical

uncertainty of the average of individual elements of the matrix generically decreases with the number of independent measurements T as $1/\sqrt{T}$. There are $N(N+1)/2$ independent elements of the correlation matrix for a system with N degrees of freedom.

Consider a statistical system consisting of N real degrees of freedom $x_i, i = 1, \ldots, N$ with a stationary probability distribution

$$p\left(x_1, \ldots, x_N\right) \prod_{n=1}^{N} dx_n \tag{5.14}$$

such that the expectation (mean) is zero:

$$\int x_i p\left(x_1, \ldots, x_N\right) \prod_{n=1}^{N} dx_n = 0, \quad \forall i \tag{5.15}$$

The covariance matrix for the system is defined as

$$C_{ij} = \int x_i x_j p\left(x_1, \ldots, x_N\right) \prod_{n=1}^{N} dx_n \tag{5.16}$$

Assume that the system belongs to the Gaussian universality class. Under this assumption, the probability distribution can be approximated by

$$p\left(x_1, \ldots, x_N\right) \prod_{n=1}^{N} dx_n = \frac{1}{\sqrt{(2\pi)^N \det \mathbf{C}}} \exp\left(-\frac{1}{2} \sum_{ij} x_i C_{ij} x_j\right) \prod_{n=1}^{N} dx_n \tag{5.17}$$

where C_{ij} is a covariance system (5.16). By construction, it is a symmetric, positive-definite matrix. In fact, for a wide class of models, the Gaussian approximation describes well the large N behavior of the system as a consequence of the central limit theorem. Deviations from the Gaussian behavior can result either from the presence of fat (heavy) tails in the probability distribution or from collective excitations of many degrees of freedom. None of these effects will be discussed here.

Experimentally, the correlation matrix is computed as follows. One performs a series of T independent measurements. Assume $T > N$. The measured vectored values $\mathbf{x}_n$ form a rectangular $N \times T$ matrix $\mathbf{X}$ with elements X_{nt}, where X_{nt} is the measured value of the n-th degree of freedom $\mathbf{x}_n$ in the t-th experiment $t = 1, \ldots, T$. The experimental correlation matrix is computed using the following estimator (called sample covariance matrix in statistics):

$$c_{ij} = \frac{1}{T} \sum_{t=1}^{T} X_{it} X_{jt} = \frac{1}{T} \left\{\mathbf{X}\mathbf{X}^T\right\}_{ij} \tag{5.18}$$

where $\mathbf{X}^T$ is the transpose of $\mathbf{X}$. We expect that for $T \to \infty$, the estimated values c_{ij} will approach the elements C_{ij}. More precisely, if the measurements are independent, the probability distribution of measuring a matrix $\mathbf{X}$ of values X_{nt} is a product of probabilities for individual measurements

$$\mathbb{P}\left(\mathbf{X}\right) D\mathbf{X} = \prod_{t=1}^{T} \left(p\left(X_{1t}, \ldots, X_{Nt}\right) \prod_{n=1}^{N} dX_{nt} \right) \tag{5.19}$$

where

$$D\mathbf{X} = \prod_{n=1}^{N} dX_{nt} \tag{5.20}$$

In particular, for the Gaussian approximation

$$\begin{aligned}\mathbb{P}(\mathbf{X})\,D\mathbf{X} &= \mathcal{N}\exp\left(-\tfrac{1}{2}\sum_{t=1}^{T}X_{it}C_{ij}^{-1}X_{jt}\right)\prod_{n,t=1}^{N,T}dX_{nt}\\ &= \mathcal{N}\exp\left(-\tfrac{1}{2}\operatorname{Tr}\mathbf{X}^{T}\mathbf{C}^{-1}\mathbf{X}\right)D\mathbf{X},\end{aligned} \tag{5.21}$$

where $\mathcal{N}$ is a normalization factor, which ensures that $\int \mathbb{P}(\mathbf{X})\,D\mathbf{X} = 1$. In this particular case, we have $\mathcal{N} = \left[(2\pi)^N \det \mathbf{C}\right]^{-T/2}$. All averages over measured values X_{nt} are calculated with this probability measure. We shall denote these averages by $\langle\cdots\rangle$, or the mathematical expectation $\mathbb{E}(\cdot)$. In particular, we see that

$$\left\langle X_{it}X_{j\tau}\right\rangle = C_{ij}\delta_{t\tau}, \qquad \mathbb{E}\left(X_{it}X_{j\tau}\right) = C_{ij}\delta_{t\tau} \tag{5.22}$$

This relation reflects the assumed absence of correlations between measurements. In general, if measurements are correlated, the right-hand side of the last equation can be expressed by a matrix $C_{it,j\tau}$ in double indices.

Now we are in a position to state the main result in this section, after the introductory remarks above. We aim to relate the true covariance matrix $\mathbf{C}$ to its estimator $\mathbf{c}$, through the use of their eigenvalue spectra. We denote the eigenvalues of the matrix $\mathbf{C}$ by $\Lambda_n, n = 1,\ldots,N$. For a given set of eigenvalues, we can calculate matrix invariants, like, for example, the spectral moments

$$M_k = \frac{1}{N}\operatorname{Tr}\mathbf{C}^k = \frac{1}{N}\sum_{n=1}^{N}\Lambda_n^k = \int d\Lambda\rho_0(\Lambda)\,\Lambda^k \tag{5.23}$$

where the density of eigenvalues $\rho_0(\Lambda)$ is defined as

$$\rho_0(\Lambda) = \frac{1}{N}\sum_{n=1}^{N}\delta\left(\Lambda - \Lambda_n\right) \tag{5.24}$$

The question is how these quantities are related to the analogous quantities defined for the estimator of the correlation matrix $\mathbf{c}$

$$m_k = \frac{1}{N}\left\langle\operatorname{Tr}\mathbf{C}^k\right\rangle = \int d\lambda\rho(\lambda)\,\lambda^k \tag{5.25}$$

where the eigenvalue density of the matrix estimator is

$$\rho(\lambda) = \frac{1}{N}\left\langle\sum_{n=1}^{N}\delta\left(\lambda - \lambda_n\right)\right\rangle \tag{5.26}$$

We expect that the dependence of the estimated spectrum $\rho(\lambda)$ and the genuine spectrum $\rho_0(\Lambda)$ should be controlled by T and N. Indeed, as we shall see later, it turns out that for $N \to \infty$, this dependence is governed by the parameter $c = N/T$, which we assume to be finite.

In order to derive the relation between the spectral properties of the covariance matrix and its estimator, it is convenient to define resolvents (Green's functions)

$$\mathbf{G}(Z) = \left(Z\mathbf{I}_N - \mathbf{C}\right)^{-1} \tag{5.27}$$

and

$$\mathbf{g}(z) = \left\langle\left(z\mathbf{I}_N - \mathbf{C}\right)^{-1}\right\rangle = \left\langle\left(z\mathbf{I}_N - \frac{1}{T}\mathbf{X}\mathbf{X}^{T}\right)^{-1}\right\rangle \tag{5.28}$$

where Z and z are complex variables. The symbol $\mathbf{I}_N$ stands for the $N \times N$ unit matrix. Expanding the resolvents in $1/Z$ (or $1/z$) in power series, we see that they can be interpreted as generating functions for the moments

$$M(Z) = \frac{1}{N} \operatorname{Tr} [Z\mathbf{G}(Z)] - 1 = \sum_{k=1}^{\infty} \frac{1}{Z^k} M_k \tag{5.29}$$

and

$$m(z) = \frac{1}{N} \operatorname{Tr} \left[z\mathbf{g}(z)\right] - 1 = \sum_{k=1}^{\infty} \frac{1}{z^k} m_k \tag{5.30}$$

From the relation between $M(Z)$ an $m(z)$, we can determine the corresponding relation between the eigenvalue spectra $\rho_0(\Lambda)$ and $\rho(\lambda)$. Indeed, taking the imaginary part of $\frac{1}{N} \operatorname{Tr} \mathbf{g}(z)$ (and $\frac{1}{N} \operatorname{Tr} \mathbf{G}(Z)$) for $z = \lambda + i0^+$ (or $Z = \Lambda + i0^+$), where λ and Λ are real, we can calculate the eigenvalue densities $\rho(\lambda)$ (and $\rho_0(\Lambda)$) directly:

$$\rho(\lambda) = -\frac{1}{\pi} \operatorname{Im} \frac{1}{N} \operatorname{Tr} \mathbf{g}\left(\lambda + i0^+\right) \tag{5.31}$$

as follows from the standard relation for distributions:

$$\frac{1}{x + i0^+} = PV\frac{1}{x} - i\pi\delta(x)$$

where PV stands for principal value and x is real.

The fundamental relation between the generating functions (5.29) and (5.30) is derived by means of a diagrammatic technique [123] for calculating integrals (5.28) with the Gaussian measure (5.21). The large N limit corresponds to the planar limit in which only planar diagrams contribute. This significantly simplifies considerations and allows us to write down closed formulae for the resolvents.

The *fundamental relation* between the generating functions (5.29) and (5.30) reads

$$m(z) = M(Z) \tag{5.32}$$

where the complex number Z is related to z by the conformal map

$$Z = \frac{z}{1 + cm(z)} \tag{5.33}$$

or, equivalently, if we invert the last relation for $z = z(Z)$, as

$$z = Z(1 + cM(Z)) \tag{5.34}$$

We can use (5.32) and (5.33) to compute moments of the genuine correlation function $\mathbf{C}$ from the experimentally measured moments of the estimator $\mathbf{c}$. Indeed, combining (5.32) and (5.33), we obtain the following equation:

$$m(z) = M\left(\frac{z}{1 + cm(z)}\right) \tag{5.35}$$

which gives a compact relation between moments m_k and M_k:

$$\sum_{k=1}^{\infty} \frac{m_k}{z^k} = \sum_{k=1}^{\infty} \frac{M_k}{z^k} \left(1 + c \sum_{l=1}^{\infty} \frac{m_l}{z^l}\right)^k \tag{5.36}$$

from which we can recursively express m_k by $M_l, l = 1, \ldots, k$

$$\begin{aligned} m_1 &= M_1 \\ m_2 &= M_2 + cM_1^2 \\ m_3 &= M_3 + 3cM_1M_2 + c^2M_1^3 \\ &\cdots \end{aligned} \tag{5.37}$$

or inversely: M_k by $m_l, l = 1, \ldots, k$

$$\begin{aligned} M_1 &= m_1 \\ M_2 &= m_2 - cm_1^2 \\ M_3 &= m_3 - 3cm_1m_2 + 2c^2m_1^3 \\ &\cdots \end{aligned} \tag{5.38}$$

Let us observe that for $c < 1$ the functions $M(Z)$ an $m(z)$ (expressed in infinite power series) can also be expanded around $z = Z = 0$. In this case

$$M(Z) = -\sum_{k=0}^{\infty} Z^k M_{-k}$$

where

$$M_{-k} = \frac{1}{N} \operatorname{Tr} \mathbf{C}^{-k}$$

Similarly

$$m(Z) = -\sum_{k=0}^{\infty} z^k m_{-k}$$

where

$$m_{-k} = \frac{1}{N} \left\langle \operatorname{Tr} \mathbf{c}^{-k} \right\rangle$$

Using the same manipulation as before, we obtain

$$\sum_{k=1}^{\infty} M_{-k} Z^k = \sum_{k=1}^{\infty} m_{-k} Z^k \left(1 - c - c \sum_{l=1}^{\infty} M_{-l} Z^l \right)^k \tag{5.39}$$

and hence

$$\begin{aligned} M_{-1} &= (1-c)\, m_{-1} \\ M_{-2} &= (1-c)^2 m_{-2} - c\,(1-c)\, m_{-1}^2 \\ M_{-3} &= (1-c)^3 m_{-3} - c(1-c)^2 m_{-1} m_{-2} - c^2\,(1-c)\, m_{-1}^3 \\ &\cdots . \end{aligned}$$

The relations between moments can be used directly in practical applications to clean the spectrum of the correlation matrix. Formulae (5.32) and (5.33) encode full information about the relation between the eigenvalue spectrum $\rho_0(\Lambda)$ and $\rho(\lambda)$ for a given c. In particular, if one knows the spectrum $\rho_0(\Lambda)$ of the correlation matrix $\mathbf{C}$, we can exactly determine, for a given c, the shape of the spectrum $\rho(\lambda)$ of the estimator dressed by statistical fluctuations.

The algorithm is summarized as follows:

- From the eigenvalue spectrum $\rho_0(\Lambda)$, we first deduce an explicit form of the function $M(Z)$ and of the right-hand side of (5.34).
- Inverting (5.34) for Z, we find the dependence $Z = Z(z)$ as a function of z.
- Inserting it to (5.32), we determine the function $m(z)$.
- Taking the imaginary part along the cuts of the map $m(z)$ on the real axis (5.31), we eventually find $\rho(\lambda)$.

One can easily write a numerical program that realizes this procedure. In few cases the solution is possible analytically.

Example 5.2.1 (correlation matrix C with degenerate eigenvalues) Consider the correlation matrix **C** whose spectrum is given by a sequence of degenerate eigenvalues $\mu_i, i = 1, \ldots, K$ with degeneracies n_i. Consequently, defining $p_i = n_i/N$, $\sum_{i=1}^{K} p_i = 1$, we have

$$M(Z) = \sum_{i=1}^{K} \frac{p_i \mu_i}{Z - \mu_i} \tag{5.40}$$

This form is particularly simple to discuss. We should, however, keep in mind that relations (5.32) and (5.33) also remain valid in a more general case, for instance, when in the limit $N \to \infty$, the spectrum of $\rho_0(\Lambda)$ is not a sum of delta functions but approaches some continuous distribution. Map (5.34) now reads

$$z = Z\left(1 + c\sum_{i=1}^{K} \frac{p_i \mu_i}{Z - \mu_i}\right) \tag{5.41}$$

Clearly, if we solve this equation for $Z = Z(z)$, we obtain a multivalued function, except in the case $c = 0$ for which we have a simple relation, $z = Z$. The "physical" Riemann sheet of the map $Z = Z(z)$ is singled out by the condition $Z \to z$ for $z \to \infty$. On this sheet, the complex z-plane is mapped on a part of the Z plane without a simply or multiply connected region surrounding the poles at $Z = \mu_i$. □

Example 5.2.2 (correlation matrix C with one eigenvalue) Let us consider the simplest case, as an illustration, where $K = 1$. In this case, we have only one eigenvalue $\mu_1 = \mu$ and $p_1 = 1$ and $M(Z) = \mu/(Z - \mu)$. Map (5.33) has the following form:

$$z = Z + c\frac{Z\mu}{Z - \mu}$$

If one rewrites the right-hand side of this equation using polar coordinates (R, ϕ) around the pole: $Z - \mu = \mathrm{R}e^{i\phi}$:

$$z = \mathrm{R}e^{i\phi} + c\frac{\mu^2}{R}e^{-i\phi} + \mu(1 + c)$$

one can see that the equation is invariant under the "duality" transformation

$$R \leftrightarrow c\frac{\mu^2}{R}, \qquad \phi \leftrightarrow -\phi$$

which maps the inside of the circle $|Z-\mu| = \mu\sqrt{c}$ onto the outside and vice versa. Obviously, the outside corresponds to the "physical" Riemann sheet of the inverse map $Z = Z(z)$

$$Z = \frac{1}{2}\left[(1-c)\,\mu + z + \sqrt{(z-\mu_+)\,(z-\mu_-)}\right]$$

since in this region $Z \sim z$ for $z \to \infty$. The two constants in the last equation are $\mu_\pm = \left(1 \pm \sqrt{c}\right)^2 \mu$. Along the cut $\mu_- < z < \mu_+$ on the real axis the map $Z = Z(z)$ becomes complex and ambiguous: it has a phase (sign) ambiguity, which is related to the fact that the cut is mapped into the limiting circle where the two Riemann sheets meet.

From (5.32), we can easily find the generating function $m(z)$ and then from (5.31), the spectral density of the correlation matrix $\mathbf{c}$

$$\rho(\lambda) = \frac{1}{2\pi c\mu}\,\frac{\sqrt{(\mu_+ - \lambda)\,(\lambda - \mu_-)}}{\lambda}, \tag{5.42}$$

This is a well known result called Marchenko–Pastur distribution in random matrix theory for the spectral distribution of the Wishart ensemble. It is interesting to interpret this result as a statistical smearing of the initial spectral density $\rho_0(\Lambda)$, given by the delta function localized at μ into a wide peak $\rho(\lambda)$ supported by the cut $[\mu_-, \mu_+]$, due to a finite series of measurements. The larger c, the larger is the width of the resulting distribution $\rho(\lambda)$. For instance, this formula gives $c_r \approx 0.01$ if the correlation matrix $\mathbf{C}$ has two eigenvalues $\mu_1 = 1$ and $\mu_2 = 1.1$ and the corresponding weights $p_1 = p_2 = 1/2$. □

Example 5.2.3 (correlation matrix C with two eigenvalues) The genuine covariance matrix $\mathbf{C}$ has two different eigenvalues μ_1, μ_2 with relative weights $p_1, p_2, p_1 + p_2 = 1$. In this one can also find an explicit form of the map $Z(z)$ solving the corresponding cubic (Cardano) equation.

Depending on the parameters μ_i, p_i, the map $Z(z)$ has one or two cuts on the real axis of the z plane. This means that corresponding eigenvalue distribution $\rho(\lambda)$ has a support on one or two intervals. The critical value at which a single cut solution splits into a two-cut one is:

$$c_r = \frac{\left(\mu_2 - \mu_1\right)^2}{\left[\left(p_1\mu_1^2\right)^{1/3} + \left(p_2\mu_2^2\right)^{1/3}\right]^3} \tag{5.43}$$

Thus, in this case, to observe a bimodal signal in the measured spectrum one has to perform T measurements with T of order $100N$. In Figure 5.1, we illustrate (5.43). □

The method can be straightforwardly generalized from $K = 1, 2$ to arbitrary K, $\mu_1, \ldots, \mu_K$ with $\sum_{k=1}^{K} p_i = 1$, although only the $K = 3$ case is solvable analytically (quartic Ferrari equation). In other cases, one can use a numerical implementation of the general procedure, which we described before, to determine the shape of the spectrum of the estimator $\rho(\lambda)$ from any given distribution $\rho_0(\Lambda)$ and for any c.

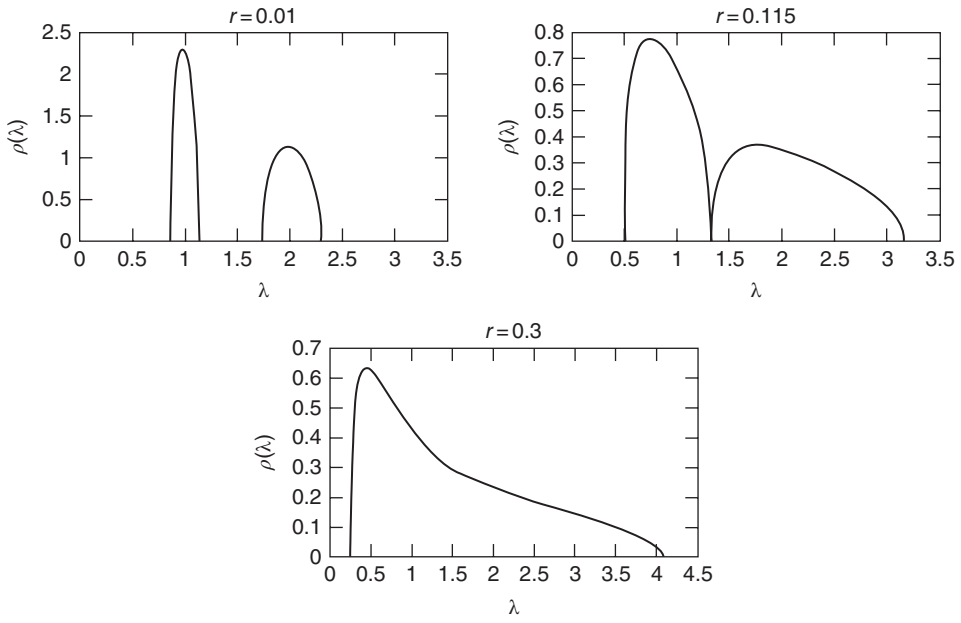


Figure 5.1 The figures represent spectra of the eigenvalue distributions $\rho(\lambda)$ of the experimental correlation matrix measured in a series of measurements for $c = r = 0.01, 0.115, 0.3$, respectively. The underlying correlation matrix has two eigenvalues $\mu_1 = 1$ and $\mu_2 = 2$ with the weights $p_1 = p_2 = 0.5$. At the critical value $c = c_r = r_c = 0.115$ ((5.43)), the spectrum splits. The spectral densities are calculated analytically. Source: Reproduced with permission from [121].

In practice, one is, however, interested in the opposite problem, that is, in the determination of the spectrum $\rho_0(\Lambda)$ of the genuine correlation matrix $\mathbf{C}$ from the distribution of the measured eigenvalues.

5.3 Wishart-Levy Free Stable Random Matrices

Consider $i = 1, \ldots, N$ stochastic time series x_{ij} observed at synchronous times $t_j, j = 0, \ldots, T$. The data can be arranged in a $N \times T$ matrix $\mathbf{M}$ of increments $m_{ij} = x_{ij} - x_{i,j-1}$, where each row corresponds to a time series and each column to a sampling time. Assuming that the average of the increments is zero, the Pearson estimator for the covariance of two time series i and j is

$$c_{ij} = \frac{1}{T} \sum_{k=1}^{T} m_{ik} m_{jk} \tag{5.44}$$

The covariances of all pairs can be collected in a $N \times N$ symmetric matrix

$$\mathbf{C} = \frac{1}{T} \mathbf{M}\mathbf{M}^T \tag{5.45}$$

The covariance matrix $\mathbf{C}$ is also called Wishart matrix as it was studied by him [116]. One is often interested in testing the hypothesis that there are no significant correlations. This can be done by comparing the eigenvalue spectrum of an empirical correlation

matrix with the spectrum of a reference matrix built with synthetic uncorrelated time series. If the matrix rows are random walks whose increments are independent and identically distributed (i.i.d.) normal deviates with standard deviation σ, the spectrum describing the above null hypothesis in the limit for $N, T \to \infty$ with $c = N/T$ is given analytically by the Marchenko–Pastur law [172]:

$$\begin{aligned} \rho_{\mathbf{C}}(\lambda) &= \frac{1}{2\pi\sigma^2 c\lambda}\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)} \\ \lambda_\pm &= \sigma^2\left(1 \mp \sqrt{c}\right)^2 \end{aligned} \tag{5.46}$$

Indeed, for a sufficiently large matrix, the exact distribution of its elements becomes less and less relevant, and the Marchenko–Pastur law can be obtained for i.i.d. increments drawn from any symmetric distribution with a finite second moment σ^2. This effect was also evident in Wigner's studies of matrices whose elements are binary random variables assuming the values ± 1 with equal probability. In both the Wigner and Wishart ensembles, the spectra of large matrices converge to that of an infinite matrix (respectively the semicircle law and the Marchenko–Pastur law) as a consequence of a generalized central limit theorem.

A practical use of (5.46) is that if the empirical spectrum of data shows significant differences from the theoretical curve, then it may be justified to reject the null hypothesis of no true correlations. The details of the latter are then a separate issue. In principle, it is possibile to test not only correlation, but also any kind of suitable assumption leading to a given shape of the expected spectrum, both theoretically and numerically. Depending on the specific case, one chooses a suitable null hypothesis.

The result given by (5.46) lies within classical random matrix theory and requires i.i.d. matrix elements with finite moments. In this section, we are concerned with the Wishart–Levy ensemble as a natural extension of the Wishart–Gaussian ensemble treated by the Marchenko–Pastur theory. The situation becomes more complicated if the elements of $\mathbf{M}$ are distributed with power-law tails, as happens in numerous physical, biological and economic data [256].

The Marchenko–Pastur theory is not valid any more when the second moment is not finite, and the corresponding spectral densities cannot be obtained from a simple extension of Gaussian random matrix theory. As a consequence of the central limit theorem for scale-free processes, the distribution of many of the above phenomena is usually assumed to be a symmetric Levy α-stable distribution, whose pdf is given most suitably as the inverse Fourier (cosine) transform of its characteristic function:

$$L_\alpha(x) = \mathcal{F}^{-1}\left[e^{-|\gamma\omega|^\alpha}\right](x) = \frac{1}{\pi}\int_0^\infty e^{-(\gamma\omega)^\alpha}\cos(x\omega)\,d\omega \tag{5.47}$$

The second and higher moments of $L_\alpha(x)$ diverge for $\alpha < 2$, and for $\alpha \leq 1$ even the first moment does not exist. If $\alpha = 2$, (5.47) gives a Gaussian with standard deviation $\sigma = \sqrt{2}\gamma$. However, we shall see that the functional representation of this distribution is not required in the derivation of the spectrum.

A matrix whose elements are i.i.d. samples from a stable density is called a Levy matrix. A symmetric Levy matrix is called a Wigner–Levy matrix. A symmetric matrix $\mathbf{C}$ built from a Levy matrix $\mathbf{M}$ according to the equation

$$\mathbf{C} = \frac{1}{T^{2/\alpha}}\mathbf{M}\mathbf{M}^T \tag{5.48}$$

is called a Wishart-Levy matrix. Notice that the normalization factor has been generalized with respect to (5.48) to take into account Levy α-stable statistics. Sampling the elements from the probability density function

$$f_X(x) = N^{2/\alpha} L_\alpha\left(N^{2/\alpha} x\right) \tag{5.49}$$

the limiting spectrum becomes independent of the matrix size N. The spectra of these matrices no longer have a finite support as in the semicircle and Marchenko–Pastur laws and are dominated by the behavior of the power-law tail of $L_\alpha(x)$.

The theory of free probability with its convenient machinery leading to analytic results that could be obtained otherwise only by means of a painful use of combinatorics. A free Levy stable random matrix has a spectrum belonging to the class of free stable laws.

5.4 Basic Concepts for Free Random Variables

We have the following exact formula:

free probability theory= noncommutative probability theory + free independence.

A symmetric $N \times N$ matrix $\mathbf{X}$ has real eigenvalues $\lambda_1, \ldots, \lambda_N$. The spectral density of $\mathbf{X}$ can be written as

$$\rho_{\mathbf{X}}(\lambda) = \frac{1}{N} \sum_{i=1}^{N} \delta\left(\lambda - \lambda_i\right) \tag{5.50}$$

where it is assumed that the weight of each eigenvalue is the same and each eigenvalue is counted as many times as its multiplicity. The resolvent matrix [257] is defined as

$$\mathbf{G}_{\mathbf{X}}(z) = \left(z\mathbf{I}_N - \mathbf{X}\right)^{-1}, \quad z \in \mathbb{C} \tag{5.51}$$

where $\mathbf{I}_N$ is the $N \times N$ identity matrix. The Green's function is defined as

$$G_{\mathbf{X}}(z) = \frac{1}{N} \operatorname{Tr} \mathbf{G}_{\mathbf{X}}(z) \tag{5.52}$$

where the trace Tr of a square matrix is defined as the sum of its diagonal elements. If $\mathbf{X}$ is a random matrix, the above definition is generalized including an expectation operator denoted by $\mathbb{E}$ (or $\langle \cdots \rangle$)

$$G_{\mathbf{X}}(z) = \frac{1}{N} \mathbb{E}\left[\operatorname{Tr} \mathbf{G}_{\mathbf{X}}(z)\right], \qquad G_{\mathbf{X}}(z) = \frac{1}{N} \langle \operatorname{Tr} \mathbf{G}_{\mathbf{X}}(z) \rangle \tag{5.53}$$

The Green's function contains the same information as the eigenvalues and the eigenvalue density of $\mathbf{X}$ [258]. The Green's function can be written in terms of the eigenvalues of $\mathbf{X}$:

$$G_{\mathbf{X}}(z) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{z - \lambda_i} \tag{5.54}$$

This is a special case of the definition through the Cauchy transform (Stieltjes transform) of a generic spectral density:

$$G_{\mathbf{X}}(z) = \int_{-\infty}^{+\infty} \frac{1}{x - \lambda} \rho_{\mathbf{X}}(\lambda)\, d\lambda \tag{5.55}$$

By using the following representation of Dirac's δ-function

$$\frac{1}{x \pm i\varepsilon} = \mathrm{PV}\left(\frac{1}{x}\right) \mp i\pi\delta(x) \tag{5.56}$$

where PV denotes the principal value and x is real, the spectral density can be obtained from the Green's function:

$$\rho_{\mathbf{X}}(\lambda) = \lim_{\varepsilon \to 0+} \frac{1}{\pi} \mathrm{Im}\left[G_{\mathbf{X}}(\lambda - i\varepsilon)\right] \tag{5.57}$$

This means that the eigenvalues follow from the discontinuities of $G_{\mathbf{X}}(z)$ on the real axis.

Noncommutativity of matrices and, in general, of operators makes it difficult to extend standard probability theory to matrices as well as operator spaces. Among possible extensions of probability theory to operator spaces, the so-called free probability theory has the advantage that many results can be deduced from well known theorems on analytic functions [125].

Conventional Classical Probability

In order to explain the framework of free probability, let us start from conventional classical probability. A probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a measure space, where Ω is the sample space, $\mathcal{F}$ is a σ-algebra on Ω, and

$$\mathbb{P} : \mathcal{F} \to [0,1] \in \mathbb{R}$$

is a non-negative measure on sets in $\mathcal{F}$ obeying Kolmogorov's axioms; $\omega \in \Omega$ is called an elementary event, $A \in \mathcal{F}$ is called an event.

A random variable $X : \Omega \to \mathbb{R}$ is a measurable function that maps elements from the sample space to the real numbers, and thus elements from $\mathcal{F}$ to a Borel σ-algebra Σ on $\mathbb{R}$. The probability distribution of X with respect to $\mathbb{P}$ is described by a measure μ_X on $(\mathbb{R}, \Sigma)$ defined as the image measure of

$$\mu_X = \mathbb{P}\left[X^{-1}(B)\right]$$

where B is any Borel set and $X^{-1}(B) \subset \mathcal{F}$ is the counter image of B. The cumulative distribution function of X is

$$F_X(x) = \mu_X(X \leqslant x).$$

The expectation value for any bounded Borel function $g : \mathbb{R} \to \mathbb{R}$ is

$$\mathbb{E}\left[g(X)\right] = \int_{\mathbb{R}} g(x)\mu_X(dx) = \int_{\mathbb{R}} g(x)dF_X(x) \tag{5.58}$$

If $F_X(x)$ is differentiable, the probability density function (pdf) of X is $f_X(x) = dF_X(x)/dx$.

This construction can be extended to noncommutative variables, such as matrices or more general operators.

Non-commutative Variables

Let $\mathcal{A}$ denote a unital algebra over a field $\mathcal{F}$, i.e. a vector space equipped with a bilinear product

$$\circ : \mathcal{A} \times \mathcal{A} \to \mathcal{A}$$

that has an identity element $\mathbf{I}$. A tracial state on $\mathcal{A}$ is a positive linear function

$$\tau : \mathcal{A} \to \mathbb{F}$$

with the properties

$$\tau(\mathbf{I}) = 1 \qquad \text{and} \qquad \tau(\mathbf{XY}) = \tau(\mathbf{YX})$$

for every $\mathbf{X}, \mathbf{Y} \in \mathcal{A}$. The couple $(\mathcal{A}, \tau)$ is called a noncommutative probability space.

For our purposes $\mathcal{A} = \mathcal{B}(\mathcal{H})$, where $\mathcal{B}(\mathcal{H})$ denotes the Banach algebra of linear operators on a real separable Hilbert space $\mathcal{H}$. This is a $*$-algebra, as it is equipped with an involution (the adjoint operation)

$$\mathbf{X} \mapsto \mathbf{X}^* : \mathcal{B}(\mathcal{H}) \to \mathcal{B}(\mathcal{H})$$

Considering a self-adjoint operator $\mathbf{X} \in \mathcal{B}(\mathcal{H})$, it is possible to associate a (spectral) distribution to $\mathbf{X}$ as in classical probability. Thanks to the Riesz representation theorem and the Stone–Weierstrass theorem, there is a unique measure $\mu_{\mathbf{X}}$ on $(\mathbb{R}, \Sigma)$ satisfying

$$\int_{\mathbb{R}} g(x)\mu_{\mathbf{X}}(dx) = \tau\left[g(\mathbf{X})\right] \tag{5.59}$$

where $g : \mathbb{R} \to \mathbb{R}$ is any bounded Borel function [259]. Therefore, we say that the distribution of $\mathbf{X}$ is described by the measure $\mu_{\mathbf{X}}$. For our purposes, this measure is equal to the spectral density $\rho_{\mathbf{X}}$ defined in (5.57). In random matrix theory, the Wigner semicircle law has the role of the Gaussian law in classical probability, and the Marchenko–Pastur law corresponds to the χ^2 law.

Independence and Freeness

Classically, independence between two random variables X and Y can be defined requiring that for any couple of bounded Borel functions f, g

$$\mathbb{E}\left[\left(f(X) - \mathbb{E}\left[f(X)\right]\right)\left(g(Y) - \mathbb{E}\left[g(Y)\right]\right)\right] = 0 \tag{5.60}$$

Analogously, two elements $\mathbf{X}$ and $\mathbf{Y}$ in a noncommutative probability space are defined as free (or freely) independent with respect to τ, if for any couple of bounded Borel functions f, g

$$\tau\left[\left(f(X) - \tau\left[f(X)\right]\right)\left(g(Y) - \tau\left[g(Y)\right]\right)\right] = 0 \tag{5.61}$$

Defining freeness between more than two elements is a nontrivial extension [260].

Generally, square $N \times N$ random matrices $\mathbf{X}$ are noncommutative random variables with respect to the function

$$\tau(\mathbf{X}) = \frac{1}{N}\mathbb{E}\left[\mathrm{Tr}\,\mathbf{X}\right]$$

(see (5.53)), but for any given N, no pair of random matrices is free. Two random matrices $\mathbf{X}, \mathbf{Y}$ can, nevertheless, reach freeness asymptotically if, for any integer $n > 0$ and any set of non-negative integers $(\gamma_1, \ldots, \gamma_n)$ and $(\beta_1, \ldots, \beta_n)$ for which in the limit $N \to \infty$

$$\tau(\mathbf{X}^{\gamma_1}) = \cdots = \tau(\mathbf{X}^{\gamma_n}) = \tau\left(\mathbf{Y}^{\beta_1}\right) = \cdots = \tau\left(\mathbf{Y}^{\beta_n}\right) = 0 \tag{5.62}$$

we have

$$\tau\left(\mathbf{X}^{\gamma_1}\mathbf{Y}^{\beta_1}\cdots\mathbf{X}^{\gamma_n}\mathbf{Y}^{\beta_n}\right) = 0 \tag{5.63}$$

This means that *large random matrices can be good approximations of free noncommutative variables*. This observation is the foundation for exploiting free noncommutative variables to handle large random matrices that naturally represent the big data.

Properties

Given an operator $\mathbf{X} \in \mathcal{B}(\mathcal{H})$, the following functions are useful in deriving its spectral distribution $\mu_{\mathbf{X}}$:

- *Moment generating function*, defined as

$$M_{\mathbf{X}}(z) = zG_{\mathbf{X}}(z) - 1 \tag{5.64}$$

The name stems from the fact that, if the distribution of $\mathbf{X}$ has finite moments of order k, $m_{\mathbf{X},k} = \tau\left(\mathbf{X}^k\right)$

$$M_{\mathbf{X}}(z) = \sum_{k=1}^{\infty} \frac{m_{\mathbf{X},k}}{z^k} \tag{5.65}$$

This can be seen inserting the sum of the geometric series

$$\sum_{k=0}^{\infty} q^k = \frac{1}{1-q}, \quad |q| < 1 \tag{5.66}$$

with $q = \lambda/|z|$ into (5.55):

$$\begin{aligned} G_{\mathbf{X}}(z) &= \int_{-\infty}^{+\infty} \frac{1}{z(1-\lambda/z)} \rho_{\mathbf{X}}(\lambda)\, d\lambda \\ &= \int_{-\infty}^{+\infty} \frac{1}{z} \sum_{k=0}^{\infty} \frac{\lambda^k}{z^k} \rho_{\mathbf{X}}(\lambda)\, d\lambda \\ &= \sum_{k=0}^{\infty} \frac{1}{z^{k+1}} \int_{-\infty}^{+\infty} \lambda^k \rho_{\mathbf{X}}(\lambda)\, d\lambda \\ &= \sum_{k=0}^{\infty} \frac{m_{\mathbf{X},k}}{z^{k+1}} \end{aligned} \tag{5.67}$$

- *R-transform*. In classical probability the pdf of the sum of two independent random variables $X + Y$ is equal to the convolution of the individual pdfs:

$$f_{X+Y}(x) = \left(f_X * f_Y\right)(x) \tag{5.68}$$

The convolution is done conveniently in Fourier space, where it becomes a multiplication: the characteristic function

$$\hat{f}_{X+Y}(\omega) = \int_{\mathbb{R}} f_{X+Y}(x)\, e^{i\omega x} \tag{5.69}$$

of $X + Y$ is the product of the characteristic functions of X and Y

$$\hat{f}_{X+Y}(\omega) = \hat{f}_X(\omega)\hat{f}_Y(\omega) \tag{5.70}$$

and the cumulant generating function of $X + Y$ is the sum of the cumulant generating functions of X and Y:

$$\log\hat{f}_{X+Y}(\omega) = \log\hat{f}_X(\omega) + \log\hat{f}_Y(\omega) \tag{5.71}$$

The free analog of the cumulant generating function is the R-transform invented by Voiculescu [126, 259] as part of the functional inverse of the Green's function:

$$G_{\mathbf{X}}\left(R_{\mathbf{X}}(z)+\frac{1}{z}\right)=z \tag{5.72}$$

The R-transform for the sum of two free operators is the sum of their R-transforms:

$$R_{\mathbf{X}+\mathbf{Y}}(z)=R_{\mathbf{X}}(z)+R_{\mathbf{Y}}(z) \tag{5.73}$$

The free analogue of convolution is indicated with the symbol $\boxplus$

$$\mu_{\mathbf{X}\boxplus\mathbf{Y}}=\mu_{\mathbf{X}}\boxplus\mu_{\mathbf{Y}} \tag{5.74}$$

This is computed through $R_{\mathbf{X}}$, given the connection between the Green's function $G_{\mathbf{X}}$ and the spectral distribution $\mu_{\mathbf{X}}$.

- *Blue function.* It is convenient to introduce also an inverse of the Green's function $G_{\mathbf{X}}$, called Blue function:

$$G_{\mathbf{X}}\left(B_{\mathbf{X}}(z)\right)=B_{\mathbf{X}}\left(G_{\mathbf{X}}(z)\right)=z \tag{5.75}$$

The Blue function is related to the R-transform by

$$B_{\mathbf{X}}(z)=R_{\mathbf{X}}(z)+\frac{1}{z} \tag{5.76}$$

- *S-transform.* In the same fashion as the R-transform for the sum, another transform allows to compute the spectral distribution of the product of two operators from their individual spectral distributions:

$$R_{\mathbf{X}}(z)=\frac{1+z}{z}\chi_{\mathbf{X}}(z) \tag{5.77}$$

where

$$\chi_{\mathbf{X}}\left(zG_{\mathbf{X}}(z)-1\right)=\frac{1}{z} \tag{5.78}$$

For $\mathbf{X}\neq\mathbf{Y}$, the S-transform of the product is the product of the individual S-transforms:

$$S_{\mathbf{XY}}(z)=S_{\mathbf{X}}(z)\,S_{\mathbf{Y}}(z) \tag{5.79}$$

As the R-transform allows to compute the free additive convolution $\boxplus$, the S-transform leads to the free multiplicative convolution $\boxtimes$:

$$\mu_{\mathbf{XY}}=\mu_{\mathbf{X}}\boxtimes\mu_{\mathbf{Y}} \tag{5.80}$$

5.5 The Analytical Spectrum of the Wishart–Levy Random Matrix

Now we are in a position to perform a case study to illustrate the machinery of free random variables.

Let $\mathbf{P}$ be the matrix projector of size $T\times T$, with N ones in arbitrary positions on the diagonal and all the other elements zero, for example:

$$\mathbf{P}=\text{diag}\,(...,1,1,\ldots,1,1,0,0,1,\ldots,1,0,\ldots) \tag{5.81}$$

Let $\mathbf{\Lambda}$ be a (large) $T \times T$ matrix with a free stable spectral distribution. This property is the analog of classical stability. The sum of two free noncommutative μ- distributed variables results in a new μ-distributed variable. The Wishart matrix ensemble of size $N \times N$ defined in (5.45) can be approximated using the $N \times T$ matrix $\mathbf{M}/T^{2/\alpha}$ obtained from $\mathbf{P\Lambda}$ if only the N nonzero rows are considered. Indicating this operation with curly braces, the approximation reads

$$\mathbf{C} = \frac{1}{T^{2/\alpha}}\mathbf{M}\mathbf{M}^T \simeq \{\mathbf{P\Lambda}\}\left\{\mathbf{\Lambda}^T\mathbf{P}\right\} \tag{5.82}$$

Our aim in this section is to find the spectrum of $\mathbf{C}$ defined in (5.82).

The moment-generating function of the $T \times T$ matrix

$$\mathbf{D} = \mathbf{\Lambda}\boldsymbol{P}\mathbf{\Lambda}^T \tag{5.83}$$

satisfies the transcendental equation

$$-\exp\left(i\frac{2\pi}{\alpha}\right) zM_{\mathbf{D}}^{2/\alpha}(z) = \left(M_{\mathbf{D}}(z)+1\right)\left(M_{\mathbf{D}}(z)+c\right) \tag{5.84}$$

which can be solved analytically for a few special values of $\alpha = 1/4, 1/3, 1/2, 2/3, 3/4, 1, 4/3, 3/2, 2$; recall that $c = N/T$ is defined above. The equation (5.84) can be solved numerically for other values.

The Green's functions of the matrices $\mathbf{D}$ and $\mathbf{C}$ are related by equation [125]:

$$G_{\mathbf{D}}(z) = c^2 G_{\mathbf{C}}(cz) + \frac{1-c}{z} \tag{5.85}$$

whence, noticing that $cG_{\mathbf{C}}(cz) = G_{\mathbf{C}}(z)$

$$M_{\mathbf{D}}(z) = zG_{\mathbf{D}}(z) - 1 = czG_{\mathbf{C}}(z) - c = cM_{\mathbf{C}}(z) \tag{5.86}$$

In the following, we will give steps that lead to (5.84) and, thus, the desired spectral density $\rho_{\mathbf{C}}(\lambda)$.

As in classical probability, stable laws have an analytic form for their Fourier transform, free stable laws have an analytic form for their Blue transform

$$B_{\mathbf{\Lambda}}(z;\alpha) = a + bz^{\alpha-1} + \frac{1}{z} \tag{5.87}$$

The parameter a accounts for a horizontal shift in the distribution of the matrix elements and can be set to zero without loss of generality. The parameter b depends on the distribution; for the symmetric Levy α-stable pdf, (5.47), b has the value [121]

$$b = e^{i\pi(\alpha/2-1)} \tag{5.88}$$

Given an index $\alpha \in (0, 2]$, $B_{\mathbf{\Lambda}}(z;\alpha)$ indirectly but precisely defines the attractor law for the sum of free variables with an α-tailed spectral distribution. Since free probability theory is exact only in the large size limit $N, T \to \infty, N/T = c$, the only variables that define the model are α and c.

Rewriting (5.87) with $G_{\mathbf{\Lambda}}(z)$ in place of z and using (5.75) yields

$$bG_{\mathbf{\Lambda}}^{\alpha-1}(z) + G_{\mathbf{\Lambda}}^{-1}(z) = z \tag{5.89}$$

which is equivalent to

$$bG_{\mathbf{\Lambda}}^{\alpha}(z) + zG_{\mathbf{\Lambda}}(z) + 1 = 0, \quad G_{\mathbf{\Lambda}}(z) \neq 0 \tag{5.90}$$

Due to (5.79) of the S-transform, if, for simplicity, from now on we substitute $\mathbf{\Lambda}$ with its symmetrized counterpart $\left(\mathbf{\Lambda}+\mathbf{\Lambda}^T\right)/2$ so that $\mathbf{\Lambda}=\mathbf{\Lambda}^T$

$$S_{\mathbf{\Lambda}\boldsymbol{P}\mathbf{\Lambda}^T}(z)=S_{\mathbf{\Lambda}}(z)\,S_{\mathbf{P}\mathbf{\Lambda}}(z)=S_{\mathbf{\Lambda}}(z)\,S_{\mathbf{\Lambda}\boldsymbol{P}}(z)=S_{\mathbf{\Lambda}\mathbf{\Lambda}\boldsymbol{P}}(z)=S_{\mathbf{\Lambda}^2\mathbf{P}}(z) \tag{5.91}$$

For the S-transform of the matrix product $\mathbf{\Lambda}^2$, we also require the Green's function. The desired relation is a consequence of the fact that the spectral measure for free Levy α-stable operators in the Wigner ensemble is symmetric [261]:

$$\rho_{\mathbf{\Lambda}}(\lambda)=\rho_{\mathbf{\Lambda}}(-\lambda),\qquad G_{\mathbf{\Lambda}}(z)=G_{-\mathbf{\Lambda}}(z) \tag{5.92}$$

We can express the Green's function of $\mathbf{\Lambda}^2$ in terms of the Green function of $\mathbf{\Lambda}$, by exploiting the Cauchy (or Stieltjes) transform representation and the previous symmetry (5.92):

$$\begin{aligned}
G_{\mathbf{\Lambda}^2}(z)&=\int_{-\infty}^{\infty}\frac{1}{z-\lambda^2}\rho_{\mathbf{\Lambda}}(\lambda)\,d\lambda\\
&=\int_{-\infty}^{\infty}\frac{1}{2\sqrt{z}}\left[\frac{1}{\sqrt{z}-\lambda}+\frac{1}{\sqrt{z}+\lambda}\right]\rho_{\mathbf{\Lambda}}(\lambda)\,d\lambda\\
&=\frac{1}{2\sqrt{z}}\left(G_{\mathbf{\Lambda}}\left(\sqrt{z}\right)+G_{-\mathbf{\Lambda}}\left(\sqrt{z}\right)\right)\\
&=\frac{1}{\sqrt{z}}G_{\mathbf{\Lambda}}\left(\sqrt{z}\right)
\end{aligned} \tag{5.93}$$

According to (5.83), the next piece in the composition of the solution is the S-transform of the projector $\mathbf{P}$, which requires its Green's function too. Inserting the spectral density of $\mathbf{P}$,

$$\rho_{\mathbf{P}}(\lambda)=c\delta(\lambda-1)+(1-c)\,\delta(\lambda), \tag{5.94}$$

into the definition of the Green's function of $\mathbf{P}$ as a Cauchy transform yields

$$\begin{aligned}
G_{\mathbf{P}}(z)&=\int_{-\infty}^{\infty}\frac{1}{z-\lambda}\rho_{\mathbf{P}}(\lambda)\,d\lambda\\
&=\int_{-\infty}^{\infty}\frac{1}{z-\lambda}\left[c\delta(\lambda-1)+(1-c)\,\delta(\lambda)\right]d\lambda\\
&=\frac{c}{z-1}+\frac{1-c}{z}
\end{aligned} \tag{5.95}$$

The moment-generating function $M_{\mathbf{P}}(z)=zG_{\mathbf{P}}(z)-1$ and the definition of the S-transform finally give

$$S_{\mathbf{P}}(z)=\frac{z+1}{z+c} \tag{5.96}$$

Rewriting (5.96) with $\sqrt{z}$ replacing z,

$$bG_{\mathbf{\Lambda}}^{\alpha}\left(\sqrt{z}\right)+\sqrt{z}G_{\mathbf{\Lambda}}\left(\sqrt{z}\right)+1=0 \tag{5.97}$$

and inserting (5.93) yields

$$bz^{\alpha/2}G_{\mathbf{\Lambda}^2}^{\alpha}(z)-zG_{\mathbf{\Lambda}^2}(z)+1=0 \tag{5.98}$$

By observing that from (5.78)

$$z=\frac{1}{\chi_{\mathbf{\Lambda}^2}\left(zG_{\mathbf{\Lambda}^2}\left(\sqrt{z}\right)-1\right)}\equiv\frac{1}{\chi_{\mathbf{\Lambda}^2}} \tag{5.99}$$

(5.98) becomes

$$\frac{b}{\chi_{\mathbf{\Lambda}^2}^{\alpha/2}} G_{\mathbf{\Lambda}^2}^{\alpha}\left(\frac{1}{\chi_{\mathbf{\Lambda}^2}}\right) - \frac{1}{\chi_{\mathbf{\Lambda}^2}} G_{\mathbf{\Lambda}^2}\left(\frac{1}{\chi_{\mathbf{\Lambda}^2}}\right) + 1 = 0 \tag{5.100}$$

Because, from (5.77), it follows that

$$\frac{1}{\chi_{\mathbf{\Lambda}^2}} G_{\mathbf{\Lambda}^2}\left(\frac{1}{\chi_{\mathbf{\Lambda}^2}}\right) - 1 = z \tag{5.101}$$

(5.100) can be simplified to

$$\frac{b}{\chi_{\mathbf{\Lambda}^2}^{\alpha/2}} G_{\mathbf{\Lambda}^2}^{\alpha}\left(\frac{1}{\chi_{\mathbf{\Lambda}^2}}\right) = z \tag{5.102}$$

Multiplying both sides by $\chi_{\mathbf{\Lambda}^2}^{-\alpha/2}/b$ yields

$$\frac{1}{\chi_{\mathbf{\Lambda}^2}^{\alpha}} G_{\mathbf{\Lambda}^2}^{\alpha}\left(\frac{1}{\chi_{\mathbf{\Lambda}^2}}\right) = \frac{z}{b}\frac{1}{\chi_{\mathbf{\Lambda}^2}^{\alpha/2}} \tag{5.103}$$

then subtracting and adding 1

$$\left[\frac{1}{\chi_{\mathbf{\Lambda}^2}} G_{\mathbf{\Lambda}^2}\left(\frac{1}{\chi_{\mathbf{\Lambda}^2}}\right) - 1 + 1\right]^{\alpha} = \frac{z}{b}\frac{1}{\chi_{\mathbf{\Lambda}^2}^{\alpha/2}} \tag{5.104}$$

and inserting (5.101) again gives

$$(z+1)^{\alpha} = \frac{z}{b}\frac{1}{\chi_{\mathbf{\Lambda}^2}^{\alpha/2}} \tag{5.105}$$

which can be written as

$$\chi_{\mathbf{\Lambda}^2} = \frac{1}{(z+1)^2}\left(\frac{z}{b}\right)^{2/\alpha} \tag{5.106}$$

Now, using (5.77), the definition of the S-transform, and the result

$$S_{\mathbf{\Lambda}^2} = \frac{1+z}{z}\chi_{\mathbf{\Lambda}^2} = \frac{1}{z(1+z)}\left(\frac{z}{b}\right)^{2/\alpha} \tag{5.107}$$

which can be used to write $S_{\mathbf{D}}$, the S-transform of the Wishart matrix on the right-hand-side of (5.82) is

$$S_{\mathbf{P}\mathbf{\Lambda}^2} = S_{\mathbf{P}} S_{\mathbf{\Lambda}^2} = \frac{1}{z(c+z)}\left(\frac{z}{b}\right)^{2/\alpha} \tag{5.108}$$

This result is the starting point for the way back. Reapplying the definition of the S-transform, we can write

$$\chi_{\mathbf{P}\mathbf{\Lambda}^2} = \frac{z}{z+1} S_{\mathbf{P}\mathbf{\Lambda}^2} = \frac{1}{(z+1)(z+c)}\left(\frac{z}{b}\right)^{2/\alpha} \tag{5.109}$$

and

$$\frac{1}{\chi_{\mathbf{P}\mathbf{\Lambda}^2}} = (z+1)(z+c)\left(\frac{z}{b}\right)^{-2/\alpha} \tag{5.110}$$

Together with $M_{\mathbf{D}}(z) = zG_{\mathbf{D}}(z) - 1$, this allows the substitution

$$\chi_{\mathbf{D}}\left(M_{\mathbf{D}}(z)\right) = 1/z, \qquad M_{\mathbf{D}}\left(1/\chi_{\mathbf{D}}\right) = z$$

Notice that we changed the index $\mathbf{\Lambda}^2\mathbf{P}$ to $\mathbf{D}$ to emphasize our goal. So we can finally write

$$z = \left(M_{\mathbf{D}}(z) + 1\right)\left(M_{\mathbf{D}}(z) + c\right)\left(\frac{M_{\mathbf{D}}(z)}{b}\right)^{-2/\alpha} \tag{5.111}$$

Inserting (5.86) yields the corresponding equation for $\mathbf{C}$

$$z = \left(cM_{\mathbf{C}}(z) + 1\right)\left(cM_{\mathbf{C}}(z) + c\right)\left(\frac{cM_{\mathbf{C}}(z)}{b}\right)^{-2/\alpha} \tag{5.112}$$

gathering c:

$$z = c^{2-2/\alpha}\left(M_{\mathbf{C}}(z) + 1/c\right)\left(M_{\mathbf{C}}(z) + 1\right)\left(\frac{M_{\mathbf{C}}(z)}{b}\right)^{-2/\alpha} \tag{5.113}$$

From (5.64) and from the relation between the moment generating function and the spectrum, we finally obtain

$$\rho_{\mathbf{C}}(\lambda) = \frac{1}{\pi\lambda}\operatorname{Im}\left[M_{\mathbf{C}}(\lambda + i0^-)\right] \tag{5.114}$$

Inserting b from (5.88) and rearranging, (5.111) takes the form anticipated in (5.84). Returning to the reason for the section, the result described by (5.113) must be considered an approximation of the curve corresponding to the null hypothesis of absence of correlation in time series with fat-tailed increments.

5.6 Basic Properties of the Stieltjes Transform

The Stieltjes transform is relevant to free random variables. We include some introductory materials for convenience.

Let G be a function of bounded variation defined on the real line. Then, its Stieltjes transform is defined by

$$m(z) \stackrel{\wedge}{=} \int_{-\infty}^{\infty} \frac{1}{x - z} G(dx) \tag{5.115}$$

where $z = u + iv$ with $v > 0$. The integrand in (5.115) is bounded by $1/v$, the integral always exists, and

$$\frac{1}{\pi}\operatorname{Im}(m(z)) = \int_{-\infty}^{\infty} \frac{v}{\pi\left[(x-u)^2 + v^2\right]} G(dx).$$

This is the convolution of G with a Cauchy density, with a scale parameter v. If G is a distribution function, then its Stieltjes transform always has a positive imaginary part. Thus, we can easily verify that, for any continuity points $x_1 < x_2$ of G,

$$\lim_{v\to 0}\int_{x_1}^{x_2} \frac{1}{\pi}\operatorname{Im}(m(z))du = G\left(x_2\right) - G\left(x_1\right) \tag{5.116}$$

(5.116) provides a continuity theorem between the family of distribution functions and the family of their Stietjes transforms.

If $\mathrm{Im}\,(m(z))$ is continuous at $x_0 + i0$, then $G(x)$ is differentiable at $x = x_0$ and its derivative equals $\frac{1}{\pi}\,\mathrm{Im}\left(m\left(x_0 + i0\right)\right)$. (5.116) gives an easy way to find the density of a distribution if its Stieltjes transform is known.

Let G be the empirical spectral distribution of a Hermitian matrix $\mathbf{A}_N$ of $N \times N$. It is seen that

$$m_G(z) = \frac{1}{N}\,\mathrm{Tr}\,(\mathbf{A} - z\mathbf{I})^{-1} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{A_{ii} - z - \boldsymbol{\alpha}_i^H\left(\mathbf{A}_i - z\mathbf{I}_{N-1}\right)^{-1}\boldsymbol{\alpha}_i} \tag{5.117}$$

where $\boldsymbol{\alpha}_i$ is the $i-$th column vector of $\mathbf{A}$ with the $i-$th entry removed and $\mathbf{A}_i$ is the matrix obtained from $\mathbf{A}$ with the $i-$th row and column deleted. (5.117) is a powerful tool in analyzing the spectrum of a large random matrix. As mentioned above, the mapping from distribution functions to their Stieltjes transforms is continuous.

Example 5.6.1 (limiting spectral distributions of the Wigner matrix) As an illustration of how to use (5.117), let us consider the Wigner matrix to find its limiting spectral distribution.

Let $m_N(z)$ be the Stieltjes transform of the empirical spectral distribution of $N^{-1/2}\mathbf{W}$. By (5.117), and noticing $w_{ii} = 0$, we have

$$\begin{aligned} m_N(z) &= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{-z - \frac{1}{N}\boldsymbol{\alpha}_i^H\left(N^{-1/2}\mathbf{W}_i - z\mathbf{I}_{N-1}\right)^{-1}\boldsymbol{\alpha}_i} \\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{-z - \sigma^2 m_N(z) + \varepsilon_i} = -\frac{1}{-z + \sigma^2 m_N(z)} + \delta_N \end{aligned}$$

where

$$\begin{aligned} \varepsilon_i &= \sigma^2 m_N(z) - \tfrac{1}{N}\boldsymbol{\alpha}_i^H\left(N^{-1/2}\mathbf{W}_i - z\mathbf{I}_{N-1}\right)^{-1}\boldsymbol{\alpha}_i \\ \delta_N &= \frac{1}{N}\sum_{i=1}^{N}\frac{-\varepsilon_i}{\left(-z - \sigma^2 m_N(z) + \varepsilon_i\right)\left(-z - \sigma^2 m_N(z)\right)} \end{aligned}$$

For any fixed $v_0 > 0$ and $B > 0$, with $z = u + iv$, we have (omitting the proof)

$$\sup_{|u| \leqslant B, v_0 \leqslant v \leqslant B}\left|\delta_N(z)\right| = o(1), \quad a.s. \tag{5.118}$$

Omitting the middle steps, we have

$$m_N(z) = -\frac{1}{2\sigma^2}\left[z + \delta_N\sigma^2 - \sqrt{\left(z - \delta_N\sigma^2\right)^2 - 4\sigma^2}\right] \tag{5.119}$$

From (5.119) and (5.118), it follows that, with probability 1, for every fixed z with $v > 0$

$$m_N(z) \to m(z) = -\frac{1}{2\sigma^2}\left[z - \sqrt{z^2 - 4\sigma^2}\right]$$

Letting $v \to 0$, we find the density of the semicircle law. □

Let $\mathbf{A}_N$ be an $N \times N$ Hermitian matrix and $F_{\mathbf{A}_N}$ be its empirical spectral distribution. If the measure μ admits a density $f(x)$ with support Ω:

$$d\mu(x) = f(x)dx \ \text{ on } \ \Omega$$

Then, the Stieltjest transform of $F_{\mathbf{A}_N}$ is given for complex arguments by

$$S_{\mathbf{A}_N}(z) = \Psi_\mu(z) = \int \frac{1}{x-z} dF_{\mathbf{A}_N}(x) = \frac{1}{N}\mathrm{Tr}\left(\mathbf{A_N} - z\mathbf{I}\right)^{-1}$$
$$= -\sum_{k=0}^{\infty} z^{-(k+1)} \left(\int_\Omega x^k f(x) dx \right) = -\sum_{k=0}^{\infty} z^{-(k+1)} M_k \tag{5.120}$$

where $M_k = \int_\Omega x^k f(x) dx$ is the $k-$th moment of F. This provides a link between the Stieltjes transform and the moments of $\mathbf{A}_N$. The moments of random Hermitian matrices become practical if direct use of the Stieltjes transform is too difficult.

Let $\mathbf{A} \in \mathbb{C}^{N\times M}, \mathbf{B} \in \mathbb{C}^{M\times N}$, such that $\mathbf{AB}$ is Hermitian. Then, for $z \in \mathbb{C}\backslash\mathbb{R}$, we have [136, p. 37]

$$\frac{M}{N} m_{F_{\mathbf{BA}}}(z) = m_{F_{\mathbf{AB}}}(z) + \frac{N-M}{N}\frac{1}{z}$$

In particular, we can apply $\mathbf{AB} = \mathbf{XX}^H$.

Let $\mathbf{X} \in \mathbb{C}^{\mathbf{N}\times\mathbf{N}}$ be Hermitian and a be a nonzero real. Then, for $z \in \mathbb{C}\backslash\mathbb{R}$

$$m_{F_{a\mathbf{X}}}(z) = \frac{1}{a} m_{F_{\mathbf{X}}}(z)$$

There are only a few kinds of random matrices for which the corresponding asymptotic eigenvalue distributions are known explicitly [262]. For a wider class of random matrices, however, explicit calculation of the moments turns out to be infeasible. The task of finding an unknown probability distribution given its moments is known as the problem of moments. It was addressed by Stieltjes in 1894 using the integral transform defined in (5.120). A simple Taylor series expansion of the kernel of the Stieltjes transform

$$-\lim_{s\to\infty} \frac{d^m}{dx^m} \frac{G\left(s^{-1}\right)}{s} = m! \int x^m dF(x)$$

shows how the moments can be found given the Stieltjes transform, without the need for integration. The probability density function can be obtained from the Stieltjes transform, simply taking the limit

$$p(x) = \lim_{y\to 0+} \frac{1}{\pi} \mathrm{Im} G(x + jy)$$

which is called the Stieltjes inverse formula [169].

We follow [263] for the following properties:

- Identical sign for imaginary part

$$\mathrm{Im}\Psi_\mu(z) = \mathrm{Im}(z) \int_\Omega \frac{f(\lambda)}{(\lambda - x)^2} d\lambda$$

  where $\Im$ is the imaginary part of $z \in \mathbb{C}$.
- Monotonicity. If $z = x \in \mathbb{R}\backslash\Omega$, then $\Psi_\mu(z)$ is well defined and

$$\Psi'_\mu(z) = \int_\Omega \frac{f(\lambda)}{(\lambda - x)^2} d\lambda > 0 \Rightarrow \Psi'_\mu(z) \nearrow \text{ on } \backslash\Omega$$

- Inverse formula

$$f(x) = \frac{1}{\pi} \lim_{y\to 0^+} \mathrm{Im}\Psi(x + jy) \tag{5.121}$$

  Note that if $x \in \mathbb{R}\backslash\Omega$, then $\Psi_\mu(x) \in \mathbb{R} \Rightarrow f(x) = 0$.

- Dirac measure. Let δ_x be the Dirac measure at x

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in \mathrm{A}, \\ 0 & else \end{cases}$$

Then,

$$\Psi_{\delta_x}(z) = \frac{1}{x-z}; \Psi_{\delta_0}(z) = -\frac{1}{z}$$

An important example is

$$L_M = \frac{1}{M}\sum_{k=1}^{M}\delta_{\lambda_k} \Rightarrow \Psi_{L_M}(z) = \frac{1}{M}\sum_{k=1}^{M}\frac{1}{\lambda_k - z}$$

- Link with the resolvent. Let $\mathbf{X}$ be a $M \times M$ Hermitian matrix

$$\mathbf{X} = \mathbf{U}\begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_M \end{pmatrix}\mathbf{U}^H$$

and consider its resolvent $\mathbf{Q}(z)$ and spectral measure L_M

$$\mathbf{Q}(z) = (\mathbf{X} - z\mathbf{I})^{-1}, L_M = \frac{1}{M}\sum_{k=1}^{M}\delta_{\lambda_k}$$

The Stieltjes transform of the spectral measure is the normalized trace of the resolvent

$$\Psi_{L_M}(z) = \frac{1}{M}\mathrm{Tr}\mathbf{Q}(z) = \frac{1}{M}\mathrm{Tr}(\mathbf{X} - z\mathbf{I})^{-1}$$

Gaussian tools [264] are useful. Let the Z_i's be independent complex Gaussian random variables and denote by $\mathbf{z} = (Z_1, \cdots, Z_n)$.

- Integration by part formula

$$\mathbb{E}\left(Z_k\Phi(\mathbf{z}, \overline{\mathbf{z}})\right) = \mathbb{E}|Z_k|^2\mathbb{E}\left(\frac{\partial\Phi}{\partial\overline{Z}_k}\right)$$

- Poincaré-Nash inequality

$$\mathrm{var}\left(\Phi(\mathbf{z}, \overline{\mathbf{z}})\right) \leq \sum_{k=1}^{n}|Z_k|^2\left(\left|\frac{\partial\Phi}{\partial Z_k}\right|^2 + \left|\frac{\partial\Phi}{\partial\overline{Z}_k}\right|^2\right)$$

5.7 Basic Theorems for the Stieltjes Transform

Theorem 5.7.1 ([265]) Let $m_F(z)$ be the Stieltjes transform of a distribution function F, then

- m_F is analytic over $\mathbb{C}^+$;
- if $z \in \mathbb{C}^+$, then $m_F(z) \in \mathbb{C}^+$;
- if $z \in \mathbb{C}^+$, $\left|m_F(z)\right| \leqslant \frac{1}{\mathrm{Im}(z)}$ and $\mathrm{Im}\left(\frac{1}{m_F(z)}\right) \leqslant -\mathrm{Im}(z)$;

- if $F(0^-) = 0$, then m_F is analytic over $\mathbb{C}\backslash\mathbb{R}^+$. Moreover, $z \in \mathbb{C}^+$ implies $zm_F(z) \in \mathbb{C}^+$ and we have the inequalities

$$\left|m_F(z)\right| \leqslant \begin{cases} \frac{1}{|\mathrm{Im}(z)|}, z \in \mathbb{C}\backslash\mathbb{R} \\ \frac{1}{|z|}, z < 0 \\ \frac{1}{\mathrm{dist}(z,\mathbb{R}^+)}, z \in \mathbb{C}\backslash\mathbb{R}^+ \end{cases}$$

with dist being the Euclidean distance.
Conversely, if $m_F(z)$ is a function analytical on $\mathbb{C}^+$ such that $m_F(z) \in \mathbb{C}^+$ if $z \in \mathbb{C}^+$ and

$$\lim_{y\to\infty} -iym_F\left(iy\right) = 1$$

then $m_F(z)$ is the Stieltjes transform of a distribution function F given by

$$F(b) - F(a) = \lim_{y\to 0} \frac{1}{\pi} \int_a^b \mathrm{Im}\left(m_F(x + jy)\right)dx.$$

If, moreover, $zm_F(z) \in \mathbb{C}^+ \text{for} z \in \mathbb{C}^+$, then $F(0^-) = 0$, in which case $m_F(z)$ has an analytic continuation on $\mathbb{C}\backslash\mathbb{R}^+$.

Our version of the above theorem is close to [136] with slightly different notation.

Let $t > 0$ and $m_F(z)$ be the Stieltjes transform of a distribution function F. Then, for $z \in \mathbb{C}^+$, we have [136]

$$\left|\frac{1}{1 + tm_F(z)}\right| \leqslant \frac{|z|}{\mathrm{Im}(z)}.$$

Let $x \in \mathbb{C}^N, t > 0$ and $\mathbf{A} \in \mathbb{C}^{N\times N}$ be Hermitian, non-negative definite. Then, for $z \in \mathbb{C}^+$ we have [136]

$$\left|\frac{1}{1 + tx^H(\mathbf{A} - z\mathbf{I})^{-1}x}\right| \leqslant \frac{|z|}{\mathrm{Im}(z)}.$$

The fundamental result in the following theorem [266] states the equivalence between pointwise convergence of the Stieltjes transform and weak convergence of probability measures.

Theorem 5.7.2 (Equivalence) Let (μ_n) be probability measures on $\mathbb{R}$ and $\left(\Psi_{\mu_n}\right)$, Ψ_{μ_n} the associated Stieltjes transform. Then the following two statements are equivalent:

- $\Psi_{\mu_n}(z) \underset{n\to\infty}{\to} \Psi_{\mu}(z)$ for all $z \in \mathbb{C}^+$
- $\mu_n \underset{n\to\infty}{\overset{w}{\to}} \mu$

Let the random matrix $\mathbf{W}$ be square $N \times N$ with i.i.d. entries with zero mean and variance $\frac{1}{N}$. Let Ω be the set containing eigenvalues of $\mathbf{W}$. The empirical distribution of the eigenvalues

$$P_{\mathbf{H}}(z) \overset{\Delta}{=} \frac{1}{N} \left|\{\lambda \in \Omega : \mathrm{Re}\lambda < \mathrm{Re}z \text{ and } \mathrm{Im}\ \lambda < \mathrm{Im}z\}\right|$$

converges a nonrandom distribution function as $N \to \infty$.

Table 5.1 Common random matrices and their moments (the entries of $\mathbf{W}$ are i.i.d. with zero mean and variance $\frac{1}{N}$; $\mathbf{W}$ is square $N \times N$, unless otherwise specified. $\mathrm{tr}\,(\mathbf{H}) \triangleq \lim_{N\to\infty} \frac{1}{N} \mathrm{Tr}\,(\mathbf{H})$).

Convergence laws	Definitions	Moments
Full-circle law	$\mathbf{W}$ square $N \times N$	
Semicircle law	$\mathbf{K} = \frac{\mathbf{W}+\mathbf{W}^H}{\sqrt{2}}$	$\mathrm{tr}\left(\mathbf{K}^{2m}\right) = \frac{1}{m+1}\begin{pmatrix} 2m \\ m \end{pmatrix}$
Quarter-circle law	$\mathbf{Q} = \sqrt{\mathbf{W}\mathbf{W}^H}$	$\mathrm{tr}\,(\mathbf{Q}^m) = \frac{2^{2m}}{\pi m} \frac{1}{\left(\frac{m}{2}+1\right)} \begin{pmatrix} m-1 \\ \frac{m-1}{2} \end{pmatrix} \forall \mathrm{m}$ odd
Deformed quarter-circle law	$\mathbf{Q}^2$ $\mathbf{R} = \sqrt{\mathbf{W}^H\mathbf{W}}$, $\mathbf{W} \in \mathbb{C}^{N\times\beta N}$	
	$\mathbf{R}^2$	$\mathrm{tr}\left(\mathbf{R}^{2m}\right) = \frac{1}{m}\sum_{i=1}^{m} \begin{pmatrix} m \\ i \end{pmatrix}\begin{pmatrix} m \\ i-1 \end{pmatrix} \beta^i$
Haar distribution	$\mathbf{T} = \mathbf{W}\left(\mathbf{W}^H\mathbf{W}\right)^{-\frac{1}{2}}$	
Inverse semicircle law	$\mathbf{Y} = \mathbf{T} + \mathbf{T}^H$	

Table 5.1 compiles some moments for commonly encountered matrices from [262]. Table 5.2 lists commonly used random matrices and their density functions. Calculating eigenvalues λ_k of a matrix $\mathbf{X}$ is not a linear operation. Calculation of the moments of the eigenvalue distribution is, however, conveniently done using a normalized trace because

$$\frac{1}{N}\sum_{k=1}^{N} \lambda_k^m = \frac{1}{N}\mathrm{Tr}\,(\mathbf{X}^m)$$

Thus, in the large matrix limit, we define $\mathrm{tr}(\mathbf{X})$ as

$$\mathrm{tr}\,(\mathbf{X}) \triangleq \lim_{N\to\infty} \frac{1}{N}\mathrm{Tr}\,(\mathbf{X})$$

Table 5.2 is self-contained and only some remarks are made here. For the Haar distribution, all eigenvalues lie on the complex unit circle because the matrix $\mathbf{T}$ is unitary. The essential nature is that the eigenvalues are uniformly distributed. The Haar distribution demands Gaussian distributed entries in the random matrix $\mathbf{W}$. This condition does not seem to be necessary, but allowing for any complex distribution with zero mean and finite variance is not sufficient.

Table 5.3[1] lists some transforms (Stieltjes, R-, S- transforms) and their properties. The Stieltjes transform is more fundamental because both R-transform and S-transform can be expressed in terms of the Stieltjes transform.

Products of Random Matrices Almost certainly, the eigenvalue distribution of the matrix product

$$\mathbf{P} = \mathbf{W}^H\mathbf{W}\mathbf{X}$$

converges in distribution, as $K, N \to \infty$ but $\beta = K/N$.

1 This table is primarily compiled from [262].

Table 5.2 Definition of commonly encountered random matrices for convergence laws (the entries of $\mathbf{W}$ are i.i.d. with zero mean and variance $\frac{1}{N}$; $\mathbf{W}$ is square $N \times N$, unless otherwise specified).

Convergence laws	Definitions	Density functions
Full-circle law	$\mathbf{W}$ square $N \times N$	$p_{\mathbf{W}}(z) = \begin{cases} \frac{1}{\pi} & \lvert z\rvert < 1 \\ 0 & elsewhere \end{cases}$
Semicircle law	$\mathbf{K} = \frac{\mathbf{W}+\mathbf{W}^H}{\sqrt{2}}$	$p_{\mathbf{K}}(z) = \begin{cases} \frac{1}{2\pi}\sqrt{4-x^2} & \lvert x\rvert < 2 \\ 0 & elsewhere \end{cases}$
Quarter-circle law	$\mathbf{Q} = \sqrt{\mathbf{W}\mathbf{W}^H}$	$p_{\mathbf{Q}}(z) = \begin{cases} \frac{1}{\pi}\sqrt{4-x^2} & 0 \le x \le 2 \\ 0 & elsewhere \end{cases}$
	$\mathbf{Q}^2$	$p_{\mathbf{Q}^2}(z) = \begin{cases} \frac{1}{2\pi}\sqrt{\frac{4-x}{x}} & 0 \le x \le 4 \\ 0 & elsewhere \end{cases}$
Deformed quarter-circle law	$\mathbf{R} = \sqrt{\mathbf{W}^H\mathbf{W}}$, $\mathbf{W} \in \mathbb{C}^{N\times\beta N}$	$p_{\mathbf{R}}(z) = \begin{cases} \frac{\sqrt{4\beta-(x^2-1-\beta)^2}}{\pi x} & a \le x \le b \\ \left(1-\sqrt{\beta}\right)^+ \delta(x) & elsewhere \end{cases}$ $a = \left\lvert 1-\sqrt{\beta}\right\rvert, b = 1+\sqrt{\beta}$
	$\mathbf{R}^2$	$p_{\mathbf{R}^2}(z) = \begin{cases} \frac{\sqrt{4\beta-(x-1-\beta)^2}}{2\pi x} & a^2 \le x \le b^2 \\ \left(1-\sqrt{\beta}\right)^+ \delta(x) & elsewhere \end{cases}$
Haar distribution	$\mathbf{T} = \mathbf{W}\left(\mathbf{W}^H\mathbf{W}\right)^{-\frac{1}{2}}$	$p_{\mathbf{T}}(z) = \frac{1}{2\pi}\delta(\lvert z\rvert - 1)$
Inverse semicircle law	$\mathbf{Y} = \mathbf{T} + \mathbf{T}^H$	$p_{\mathbf{Y}}(z) = \begin{cases} \frac{1}{\pi}\frac{1}{\sqrt{4-x^2}} & \lvert x\rvert < 2 \\ 0 & elsewhere \end{cases}$

Sums of Random Matrices Consider the limiting distribution of random Hermitian matrices of the form [172, 174]

$$\mathbf{A} + \mathbf{W}\mathbf{D}\mathbf{W}^H$$

where $\mathbf{W}(N \times K)$, $\mathbf{D}(K \times K)$, $\mathbf{A}(N \times N)$ are independent, with $\mathbf{W}$ containing i.i.d. entries having second moments, $\mathbf{D}$ is diagonal with real entries, and $\mathbf{A}$ is Hermitian. The asymptotic regime is

$$K/N \to \alpha \text{ as } \mathrm{N} \to \infty$$

The behavior is expressed using the limiting distribution function $F_{\mathbf{A}+\mathbf{W}\mathbf{D}\mathbf{W}^H}(x)$. The remarkable result is that the convergence of

$$F_{\mathbf{A}+\mathbf{W}\mathbf{D}\mathbf{W}^H}(x)$$

to a nonrandom F.

Theorem 5.7.3 ([172, 174]) Let $\mathbf{A}$ be an $N \times N$ nonrandom Hermitian matrix for which $F_{\mathbf{A}}(x)$ converge weakly as $N \to \infty$ to a distribution function $\mathbb{A}$. Let $F_{\mathbf{D}}(x)$ converge weakly as $N \to \infty$ to a distribution function denoted $\mathbb{D}$. Suppose the entries of $\sqrt{N}\mathbf{W}$

Table 5.3 Table of Stieltjes, R- and S- transforms (Table 5.2 lists the definitions of the matrix notations used in this table).

Stieltjes transform	R-transform	S-transform
$G(z) \stackrel{\Delta}{=} \int \frac{1}{x-z} dP(x)$, $\mathrm{Im} z > 0, \mathrm{Im} G(z) \geq 0$	$R(z) \stackrel{\Delta}{=} G^{-1}(-z) - z^{-1}$	$S(z) \stackrel{\Delta}{=} \frac{1+z}{z} \Upsilon^{-1}(z)$, $\Upsilon(z) \stackrel{\Delta}{=} -z^{-1} G^{-1}\left(z^{-1}\right) - 1$
$G_{\alpha \mathbf{I}}(z) = \frac{1}{\alpha - z}$	$R_{\alpha \mathbf{I}}(z) = \alpha$	$S_{\alpha \mathbf{I}}(z) = \frac{1}{\alpha}$,
$G_{\mathbf{K}}(z) = \frac{z}{2}\sqrt{1 - \frac{4}{z^2}} - \frac{z}{2}$	$R_{\mathbf{K}}(z) = z$	$S_{\mathbf{K}}(z) = \mathrm{undefined}$
$G_{\mathbf{Q}}(z) = \sqrt{1 - \frac{4}{z^2}}\left(\frac{z}{2} - \arcsin\frac{2}{z}\right) - \frac{z}{2} - \frac{1}{2\pi}$	$R_{\mathbf{Q}^2}(z) = \frac{1}{1-z}$	$S_{\mathbf{Q}^2}(z) = \frac{1}{1+z}$
$G_{\mathbf{Q}^2}(z) = \frac{1}{2}\sqrt{1 - \frac{4}{z} - \frac{1}{2}}$	$R_{\mathbf{R}^2}(z) = \frac{\beta}{1-z}$	$S_{\mathbf{R}^2}(z) = \frac{1}{\beta+z}$
$G_{\mathbf{R}^2}(z) = \sqrt{\frac{(1-\beta)^2}{4z^2} - \frac{1+\beta}{2z} + \frac{1}{4}} - \frac{1}{2} - \frac{(1-\beta)}{2z}$	$R_{\mathbf{Y}}(z) = \frac{-1+\sqrt{1+4z^2}}{z}$	$S_{\mathbf{Y}}(z) = \mathrm{undefined}$
$G_{\mathbf{Y}}(z) = \frac{-\mathrm{sign}(\mathrm{Re} z)}{\sqrt{z^2-4}}$	$R_{\alpha \mathbf{X}}(z) = \alpha R_{\mathbf{X}}(\alpha z)$	$S_{(\mathbf{Q}^2)^{-1}}(z) = S_{(\mathbf{W}^H \mathbf{W})^{-1}}(z) = -z$
$G_{\lambda^2}(z) = \frac{G_\lambda\left(\sqrt{z}\right) - G_\lambda\left(-\sqrt{z}\right)}{2\sqrt{z}}$	$\lim_{z \to \infty} R(z) = \int x dP(x)$	$S_{\mathbf{AB}}(z) = S_{\mathbf{A}}(z) S_{\mathbf{B}}(z)$
$G_{\mathbf{X}\mathbf{X}^H}(z) = \beta G_{\mathbf{X}^H \mathbf{X}}(z) + \frac{\beta - 1}{z}, \mathbf{X} \in \mathbb{C}^{N \times \beta N}$	$R_{\mathbf{A}+\mathbf{B}}(z) = R_{\mathbf{A}}(z) + R_{\mathbf{B}}(z)$	
	$G_{\mathbf{A}+\mathbf{B}}\left(R_{\mathbf{A}+\mathbf{B}}(-z) - z^{-1}\right) = z$	
$G_{\mathbf{X}^{-1}}(z) = -\frac{1}{z} - \frac{G_{\mathbf{X}}(1/z)}{2z^2}$		
$G_{(\mathbf{Q}^2)^{-1}}(z) = G_{(\mathbf{W}^H \mathbf{W})^{-1}}(z) = -\frac{1}{z} - \frac{-1+\sqrt{1-4z}}{2z^2}$		
$G_{\mathbf{X}+\mathbf{W}\mathbf{Y}\mathbf{W}^H}(z) =$ $G_{\mathbf{X}}\left(z - \beta \int \frac{y dP_{\mathbf{Y}}(x)}{1 + y G_{\mathbf{X}+\mathbf{W}\mathbf{Y}\mathbf{W}^H}(z)}\right)$		
$\mathrm{Im} z > 0, \mathbf{X}, \mathbf{Y}, \mathbf{W}$ jointly independent.		
$G_{\mathbf{W}\mathbf{W}^H}(z) = \int_0^1 u(x,z) dx,$		
$u(x,z) =$ $\left[-z + \int_0^\beta \frac{w(x,y) dy}{1 + \int_0^1 u(x',z) w(x',y) dx'}\right]^{-1},$		
$x \in [0,1]$		

i.i.d. for fixed N with unit variance (sum of the variances of the real and imaginary parts in the complex case). Then, the eigenvalue distribution of $\mathbf{A} + \mathbf{W}\mathbf{D}\mathbf{W}^H$ converges weakly to a deterministic F. Its Stieltjes transform $G(z)$ satisfies the equation:

$$G(z) = G_{\mathbb{A}}\left(z - \alpha \int \frac{\tau}{1 + \tau G(z)} d\mathbb{T}(\tau)\right)$$

Theorem 5.7.4 ([267]) Assume

- $\mathbf{X}_n = \frac{1}{\sqrt{n}}\left(X_{ij}^{(n)}\right)$, where $1 \leq i \leq n, 1 \leq j \leq p$, and $X_{i,j,N}$ are independent real random variables with a common mean and variance σ^2, satisfying

$$\frac{1}{n^2 \varepsilon_n^2} \sum_{i,j} X_{ij}^2 I\left(\left|X_{ij}\right| \geq \varepsilon_n \sqrt{n}\right) \underset{n \to \infty}{\to} 0$$

where $I(x)$ is an indication function and ϵ_n^2 is a positive sequence tending to zero.

- $\frac{p}{n} \to y > 0$ as $n \to \infty$;
- $\mathbf{T}_n$ is an $p \times p$ random symmetric matrix with $F_{\mathbf{T}_n}$ converging almost surely to a distribution $H(t)$ as $n \to \infty$;
- $\mathbf{B}_n = \mathbf{A}_n + \mathbf{X}_n \mathbf{T}_n \mathbf{X}_n^H$, where $\mathbf{A}_n$ is a random $p \times p$ symmetric matrix with $F_{\mathbf{A}_n}$ almost surely to $F_{\mathbf{A}}$, a (possibly defective) nonrandom distribution;
- $\mathbf{X}_N, \mathbf{T}_N, \mathbf{A}_N$ are independent.

Then, as $n \to \infty$, $F_{\mathbf{B}_n}$ almost certainly converges to a nonrandom distribution F, whose Stieltjes transform $m(z)$ satisfies

$$m(z) = m_{\mathbf{A}}(z)\left(z - y\int \frac{x}{1 + xm(z)} dH(x)\right)$$

Theorem 5.7.5 ([166]) Let $\mathbf{S}_n$ denote the sample covariance matrix of n pure noise vectors distributed $\mathcal{N}(0, \sigma^2 \mathbf{I}_p)$. Let l_1 be the largest eigenvalue of $\mathbf{S}_n$. In the joint limit $p, n \to \infty$, with $p/n \to c \geq 0$, the distribution of the largest eigenvalues of $\mathbf{S}_n$ converges to a Tracy–Widom distribution

$$\Pr\left\{\frac{l_1/\sigma^2 - \mu_{n,p}}{\xi_{n,p}}\right\} \to F_\beta(s)$$

with $\beta = 1$ for real valued noise and $\beta = 2$ for complex valued noise. The centering and scaling parameters, $\mu_{n,p}$ and $\xi_{n,p}$ are functions of n and p only.

Theorem 5.7.6 ([166]) Let l_1 be the largest eigenvalue as in Theorem 5.7.5. Then,

$$\Pr\left\{l_1/\sigma^2 > \left(1 + \sqrt{\frac{p}{n}}\right)^2 + \varepsilon\right\} \leq \exp\left(-nJ_{LAG}(\varepsilon)\right)$$

where

$$J_{LAG}(\varepsilon) = \int_1^x (x - y)\frac{(1+c)y + 2\sqrt{c}}{(y+B)^2}\frac{dy}{\sqrt{y^2-1}}$$

$$c = p/n, x = 1 + \frac{\varepsilon}{2\sqrt{c}}, B = \frac{1+c}{2\sqrt{c}}$$

Consider the standard model for signals with p sensors. Let $\left\{\mathbf{x}_i = \mathbf{x}(t_i)\right\}_{i=1}^n$ denote $p-$dimensional i.i.d. observations of the form

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \sigma\mathbf{n}(t) \tag{5.122}$$

sampled at n distinct times t_i, where $\mathbf{A} = \left[\mathbf{a}_1, \ldots, \mathbf{a}_K\right]^T$ is the $p \times K$ matrix of K linearly independent $p-$dimensional vectors. The $K \times 1$ vector $\mathbf{s}(t) = \left[s_1(t), \ldots, s_K(t)\right]^T$ represents the random signals, assumed zero-mean and stationary with full rank covariance matrix. σ is the unknown noise level, and $bfn(t)$ is a $p \times 1$ additive Gaussian noise vector, distributed $\mathcal{N}(0, \mathbf{I}_p)$ and independent of $\mathbf{s}(t)$.

Theorem 5.7.7 ([166]) Let $\mathbf{S}_n$ denote the sample covariance matrix of n observations from (5.122) with a single signal of strength λ. Then, in the joint limit $p, n \to \infty$, with $p/n \to c \geq 0$, the largest eigenvalue of $\mathbf{S}_n$ converges almost certainly to

$$\lambda_{\max}\left(\mathbf{S}_n\right) \overset{a.s.}{\to} \begin{cases} \sigma^2\left(1+\sqrt{p/n}\right)^2 & \lambda \leq \sigma^2\sqrt{p/n} \\ \left(\lambda+\sigma^2\right)\left(1+\frac{p}{n}\frac{\sigma^2}{\lambda}\right) & \lambda > \sigma^2\sqrt{p/n} \end{cases}$$

Theorem 5.7.8 ([268]) Let $\mathbf{C} \in {}^{p\times p}$ be positive semidefinite. Fix an integer $l \leq p$ and assume the tail

$$\left\{\lambda_i(\mathbf{C})\right\}_{i>l}$$

of the spectrum of $\mathbf{C}$ decays sufficiently fast that

$$\sum_{i>l}\lambda_i(\mathbf{C}) = \mathcal{O}\left(\lambda_1(\mathbf{C})\right)$$

Let $\left\{\mathbf{x}_i\right\}_{i=1}^n \in \mathbb{R}^p$ be i.i.d. samples drawn from a $\mathcal{N}(\mathbf{0}, \mathbf{C})$ distribution. Define the sample covariance matrix

$$\hat{\mathbf{C}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^H$$

Let κ_l be the condition number associated with a dominant $l-$dimensional invariance subspace of $\mathbf{C}$

$$\kappa_l = \frac{\lambda_1(\mathbf{C})}{\lambda_l(\mathbf{C})}$$

If

$$n = \Omega\left(\varepsilon^{-2}\kappa_l^2 l\log p\right)$$

then with high probability

$$\left|\lambda_k\left(\hat{\mathbf{C}}_n\right) - \lambda_k\left(\mathbf{C}_n\right)\right| \leq \varepsilon\lambda_k\left(\mathbf{C}_n\right), \quad \text{for } k = 1, \ldots, l$$

Theorem 5.7.8 says, assuming sufficiently fast decay of the residual eigenvalues, $n = \Omega\left(\varepsilon^{-2}\kappa_l^2 l\log p\right)$ samples ensure that the top l eigenvalues are captured with relative precision.

5.8 Free Probability for Hermitian Random Matrices

5.8.1 Random Matrix Theory

Definition 5.8.1 Consider an $n \times n$ Hermitian matrix $\mathbf{A}$ and define

$$\phi(\mathbf{A}) = \lim_{n\to\infty}\frac{1}{n}\mathrm{Tr}(\mathbf{A}) \tag{5.123}$$

The k-th moment of $\mathbf{A}$ can be expressed as $\phi\left(\mathbf{A}^k\right)$.

Theorem 5.8.2 (deformed quarter-circle law) Let the entries of the $N \times n$ matrix $\mathbf{X}$ be independent, identically distributed entries with zero mean and variance $1/N$. Then the empirical singular value distribution of $\mathbf{X}$ almost certainly converges to the limit given by

$$f_{\sqrt{\mathbf{X}\mathbf{X}^H}}(x) = \max(0, 1-c)\delta(x) + \frac{\sqrt{4c-(x^2-1-c)^2}}{\pi x}\mathbb{I}\left(\left|1-\sqrt{c}\right| \leqslant x \leqslant \left|1+\sqrt{c}\right|\right) \tag{5.124}$$

as $N, n \to \infty$ with $c = n/N$ fixed.

Moreover the transformation random variable X as $Y = X^2$ reads

$$f_Y(y) = \frac{1}{2\sqrt{y}} f_X(X)$$

This gives

$$f_{\mathbf{X}\mathbf{X}^H}(x) = \max(0, 1-c)\delta(x) + \frac{\sqrt{4c-(x-1-c)^2}}{2\pi x}\mathbb{I}\left(\left(1-\sqrt{c}\right)^2 \leqslant x \leqslant \left(1+\sqrt{c}\right)^2\right) \tag{5.125}$$

which is known as the Marchenko–Pastur distribution and its moments are given by

$$\phi\left[\left(\mathbf{X}\mathbf{X}^H\right)^K\right] = \sum_{k=1}^{K} C_{K,k} c^k. \tag{5.126}$$

On the Unitary Matrices

The $N \times N$ matrix $\mathbf{U}$ is called *unitary* if

$$\mathbf{U}^H\mathbf{U} = \mathbf{U}\mathbf{U}^H = \mathbf{I}_N \tag{5.127}$$

where $\mathbf{I}_N$ is the $N \times N$ identity matrix.

Theorem 5.8.3 (Haar distribution) Let the entries of the $N \times N$ matrix $\mathbf{X}$ be independent identically complex distributed entries with zero mean and finite positive variance. Define

$$\mathbf{U} = \mathbf{X}\left(\mathbf{X}^H\mathbf{X}\right)^{-1/2}$$

Then the empirical eigenvalue distribution of $\mathbf{U}$ converges almost surely to the limit given by

$$p_{\mathbf{U}}(z) = \frac{1}{2\pi}\delta(|z|-1)$$

as $N \to \infty$.

All the eigenvalues of the unitary Haar matrix $\mathbf{U}$ lie on the unit circle on the complex plane when the matrix size N is large.

Let the entries of the $N \times N$ matrix $\mathbf{X}$ be independent identically distributed complex entries with zero mean and finite positive variance. Then $\mathbf{X}$ can be decomposed as

$$\mathbf{X} = \mathbf{U}\mathbf{Q} \tag{5.128}$$

where $\mathbf{U}$ is a Haar matrix and $\mathbf{Q}$ fulfills the same conditions needed for the quarter-circle law defined above.

If a Hermitian random matrix $\mathbf{X}$ has the same spectral distribution with

$$\mathbf{U}\mathbf{X}\mathbf{U}^H \tag{5.129}$$

for any unitary matrix $\mathbf{U}$ independent of $\mathbf{X}$, then the matrix $\mathbf{X}$ is called unitarily invariant.

A unitarily invariant $\mathbf{X}$ can be decomposed as

$$\mathbf{U}\mathbf{\Lambda}\mathbf{U}^H$$

where $\mathbf{U}$ is a Haar matrix independent of the diagonal matrix$\mathbf{\Lambda}$.

Consider a function

$$\mathbf{Y} = g(\mathbf{X})$$

with unitarily invariant matrix $\mathbf{X}$ as an input and a Hermitian matrix $\mathbf{Y}$ as an output. Then the matrix $\mathbf{Y}$ is also *unitarily invariant*. A matrix that fulfills the same conditions needed for the semicircle law or deformed quarter-circle law, or Haar distribution is unitarily invariant.

If the joint distribution of the entries of a $N \times n$ matrix $\mathbf{X}$ is equal to the joint distribution of the entries of a matrix $\mathbf{Y}$ such that

$$\mathbf{Y} = \mathbf{U}\mathbf{X}\mathbf{V}^H \tag{5.130}$$

where $\mathbf{U}$ and $\mathbf{V}$ are Haar distributed and independent of $\mathbf{X}$, then the matrix $\mathbf{X}$ is called a *bi-unitarily invariant* random matrix.

Note that an identity matrix is also a unitary matrix. Then one can consider bi-unitarily invariant $N \times n$ random matrix $\mathbf{X}$ such that the singular value distribution of $\mathbf{X}$ is invariant both by left and right a unitary matrix product.

Let the set $\{\mathbf{X}_1, \ldots, \mathbf{X}_L\}$ consist of independent standard Gaussian matrices where the size of $\mathbf{X}_i$ is $n_i \times n_{i-1}$. Moreover, define a matrix

$$\mathbf{X} = \prod_{i=1}^{L} \mathbf{X}_i$$

Then $\mathbf{X}$ is bi-unitarily invariant.

Theorem 5.8.4 ([169]) A square random matrix $\mathbf{X}$ is a bi-unitarily-invariant, if it can be decomposed as

$$\mathbf{X} = \mathbf{U}\mathbf{Y}$$

where $\mathbf{U}$ is a Haar matrix and independent of unitarily invariant positive definite matrix $\mathbf{Y}$.

5.8.2 Free Probability Theory for Hermitian Random Matrices

Consider random matrices as noncommutative random variables in general. Then, in contrast to probability theory, we must define the variables in a matrix valued probability space or a noncommutative probability space, which changes the whole frame of (classical) probability theory.

Free probability theory is a new mathematical field that was initiated by Voiculescu in the 1980s [269]. It applies to noncommutative random variables. Large random matrices are the prime examples of free random variables.

Let $\mathcal{N}(n)$ be the set of all noncrossing permutation of $\{1, 2, \ldots, n\}$. Let π be a noncrossing partition of this set

$$\pi = \{B_1, \ldots, B_r\}$$

where B_i is the block of π that connects some elements in the (non-crossing) partition π.

Definition 5.8.5 (free cumulant) Consider a random matrix $\mathbf{X}$. Then the moment of asymptotic eigenvalue distribution can be expressed

$$\phi(\mathbf{A}^n) = \sum_{\pi \in \mathcal{N}(n)} \prod_{B_i \in \pi} \kappa_{|B_i|} \tag{5.131}$$

where κ_n is called the n-th order free cumulant.

5.8.3 Additive Free Convolution

The R-transform is the free analog of the logarithm of the Fourier transform.

Consider the free random matrices $\mathbf{A}$ and $\mathbf{B}$ and assume that their asymptotic eigenvalue distributions are known. Now we want to address how to infer the asymptotic eigenvalue distribution of $\mathbf{A} + \mathbf{B}$.

Theorem 5.8.6 (free cumulants [270]) The Hermitian matrices $\mathbf{A}$ and $\mathbf{B}$ are free. Then we have

$$\kappa_{\mathbf{A}+\mathbf{B},n} = \kappa_{\mathbf{A},n} + \kappa_{\mathbf{B},n} \tag{5.132}$$

where $\kappa_{\cdot,n}$ in (5.131).

Definition 5.8.7 (free cumulant) Consider an Hermitian random matrix $\mathbf{X}$. Then the definition of R-transform is

$$R_{\mathbf{X}}(\omega) = \sum_{n=1}^{\infty} \kappa_{\mathbf{X},n} \omega^{n-1} \tag{5.133}$$

Let the matrices $\mathbf{A}$ and $\mathbf{B}$ are free. Then, with (5.132) we have

$$\begin{aligned} R_{\mathbf{A}+\mathbf{B}}(\omega) &= \sum_{n=1}^{\infty} \left(\kappa_{\mathbf{A},n} + \kappa_{\mathbf{B},n}\right) \omega^{n-1} \\ &= \sum_{n=1}^{\infty} \kappa_{\mathbf{A},n} \omega^{n-1} + \sum_{n=1}^{\infty} \kappa_{\mathbf{B},n} \omega^{n-1} \\ &= R_{\mathbf{A}}(\omega) + R_{\mathbf{B}}(\omega) \end{aligned} \tag{5.134}$$

Example 5.8.8 (product of two i.i.d. random matrices) Let the entries of the $R \times T$ matrix $\mathbf{H}$ be i.i.d with the variance $1/R$ and the ratio $\beta = T/R$ fixed. Show that

$$R_{\mathbf{H}\mathbf{H}^H}(\omega) = \frac{\beta}{1 - \omega} \tag{5.135}$$

Our departure point is (5.126)

$$\phi\left[\left(\mathbf{X}\mathbf{X}^H\right)^n\right] = \sum_{k=1}^{n} \mathcal{N}_{n,k}\beta^k \tag{5.136}$$

With the definition of free cumulant we have

$$\sum_{k=1}^{n} \mathcal{N}_{n,k}\beta^k = \sum_{\pi\in\mathcal{N}(n)} \prod_{B_i\in\pi} \kappa_{|B_i|} \tag{5.137}$$

Note that (5.137) holds if all free cumulants are equal to β as

$$\sum_{k=1}^{n} \mathcal{N}_{n,k}\beta^k = \sum_{\pi\in\mathcal{N}(n)} \prod_{B_i\in\pi} \beta = \sum_{\pi\in\mathcal{N}(n)} \beta^r \tag{5.138}$$

Then the R-transform is given by

$$\begin{aligned} R_{\mathbf{H}^H\mathbf{H}}(\omega) &= \sum_{n=1}^{\infty} \beta\omega^{n-1} = \beta\sum_{n=1}^{\infty} \beta\omega^n \\ &= \frac{\beta}{1-\omega} \end{aligned} \tag{5.139}$$

□

Theorem 5.8.9 The functional inversion of the Stieltjes transform is equal to

$$G^{-1}(\omega) = R(\omega) + \frac{1}{\omega} \tag{5.140}$$

The R-transform of the matrix $c\mathbf{X}, c \in \mathbb{R}$ can be expressed as

$$R_{c\mathbf{X}}(\omega) = cR_{\mathbf{X}}(c\omega) \tag{5.141}$$

Example 5.8.10 (project matrix) Consider a projection matrix $\mathbf{A}$, and a matrix $\mathbf{B} = \mathbf{U}\mathbf{A}\mathbf{U}^H$, where $\mathbf{U}$ is a Haar matrix, and

$$p_{\mathbf{A}}(x) = \frac{\delta(x+1) + \delta(x-1)}{2}.$$

Find the asymptotic eigenvalue distribution of $\mathbf{A}+\mathbf{B}$. First, $\mathbf{A}$ and $\mathbf{B}$ free. So $\mathbf{A}$ and $\mathbf{B}$ have same distributions, but the eigenvectors are fully uncorrelated. Thus,

$$R_{\mathbf{A}+\mathbf{B}}(\omega) = 2R_{\mathbf{A}}(\omega) \tag{5.142}$$

The Stieltjes transform of $\mathbf{A}$ is

$$\begin{aligned} G_{\mathbf{A}}(s) &= \int \frac{1}{s-x} dP(x) \\ &= \frac{1}{2}\left(\frac{1}{s-1} + \frac{1}{s+1}\right) \end{aligned}$$

If we take the functional inverse of $G_{\mathbf{A}}(s)$, we obtain

$$\begin{aligned} \omega &= G_{\mathbf{A}}\left(G_{\mathbf{A}}^{-1}(s)\right) = G_{\mathbf{A}}\left(B_{\mathbf{A}}(s)\right) \\ &= \frac{1}{2}\left(\frac{1}{B_{\mathbf{A}}(\omega)-1} + \frac{1}{B_{\mathbf{A}}(\omega)+1}\right) \end{aligned}$$

where the Blue's function is defined as $B_{\mathbf{A}}(s) = G_{\mathbf{A}}^{-1}(s)$. Alternatively, we have

$$B_{\mathbf{A}}^2(\omega) - \frac{1}{\omega} B_{\mathbf{A}}(\omega) - 1 = 0$$

whose two solutions are

$$B_{\mathbf{A}}(\omega) = \frac{1 \mp \sqrt{1 + 4\omega^2}}{2\omega}$$

With (5.140)

$$R_{\mathbf{A}}(\omega) = B_{\mathbf{A}}(\omega) - \frac{1}{\omega} = \frac{-1 \mp \sqrt{1 + 4\omega^2}}{2\omega}$$

As from the definition of the R-transform

$$\lim_{\omega \to \infty} R_{\mathbf{A}}(\omega) = \lim_{\omega \to 0} \kappa_{\mathbf{A},1} + \sum_{k=2}^{\infty} \kappa_{\mathbf{A},k} \omega^{k-1} = \kappa_{\mathbf{A},1} = \phi(\mathbf{A})$$

where the mean is zero, we can obtain the right solution as

$$0 = \lim_{\omega \to \infty} R_{\mathbf{A}}(\omega) = \frac{-1 \mp \sqrt{1 + 4\omega^2}}{2\omega}.$$

Thus the one with positive sign is the right solution

$$R_{\mathbf{A}}(\omega) = \frac{-1 + \sqrt{1 + 4\omega^2}}{2\omega}$$

From (5.142), we have the R-transform

$$R_{\mathbf{A+B}}(\omega) = 2R_{\mathbf{A}}(\omega) = \frac{-1 + \sqrt{1 + 4\omega^2}}{\omega}$$

and the Blue's function

$$B_{\mathbf{A+B}}(\omega) = G_{\mathbf{A+B}}^{-1}(s) = \frac{\sqrt{1 + 4s^2}}{s}$$

Then, taking the inverse of the above expression, the Stieltjes transform is obtained as

$$s = \frac{\sqrt{1 + 4G_{\mathbf{A+B}}^2(s)}}{G_{\mathbf{A+B}}(s)} \Rightarrow G_{\mathbf{A+B}}(s) = \frac{1}{\sqrt{s^2 - 4}}$$

Finally, by using the inversion formula of the Stieltjes transform, we obtain the probability density function of the real eigenvalues

$$\begin{aligned} p_{\mathbf{A+B}}(x) &= -\frac{1}{\pi} \lim_{y \to 0} \Im G_{\mathbf{A+B}}(x + jy) \\ &= -\frac{1}{\pi} \lim_{y \to 0} \Im \frac{1}{\sqrt{(x + jy)^2 - 4}} \\ &= -\frac{1}{\pi} \Im \frac{1}{\sqrt{x^2 - 4}} \\ &= -\frac{1}{\pi} \frac{1}{\sqrt{\lambda^2 - 4}} \end{aligned}$$

Two observations are remarkable. First, if we randomly rotate the eigenvectors, which is a Haar-distributed operation, then the matrix $\mathbf{A}$ is free with the randomly rotated one $\mathbf{B} = \mathbf{U}\mathbf{A}\mathbf{U}^H$. Second, adding two free elements that have even discrete densities will lead to continuous density.

This example says that having an intuition about free probability in terms of (classical) probability may be false. Rather, regarding the concept of freeness as the independence of random eigenvectors is a good intuition. □

Theorem 5.8.11 (free central limit theorem [259]) Let $\mathbf{X}_k$ be a free identical family of random matrices with the eigenvalues zero mean variance 1 for all $1 \leq k \leq N$. Then the asymptotic eigenvalue distribution of

$$\mathbf{X} = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \mathbf{X}_k$$

converges in distribution to the semicircle distribution:

$$p_{\mathbf{X}}(x) = \frac{1}{2\pi}\sqrt{4 - x^2}, \quad x \in (-2, 2)$$

Proof: Using the linearity (5.134) and the scaling property (5.141) of the R-transform, the R-transform of $\mathbf{X}$ is given by

$$R_{\mathbf{X}}(\omega) = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} R_{\mathbf{X}_k}\left(\frac{\omega}{\sqrt{N}}\right)$$

As these matrices are freely identical, we have

$$\begin{aligned} R_{\mathbf{X}}(\omega) &= \frac{N}{\sqrt{N}} R_{\mathbf{X}_k}\left(\frac{\omega}{\sqrt{N}}\right) \\ &= \sqrt{N} R_{\mathbf{X}_k}\left(\frac{\omega}{\sqrt{N}}\right) \\ &= \sqrt{N}\left(\kappa_1 + \kappa_2 \frac{\omega}{\sqrt{N}} + \kappa_3 \frac{\omega^2}{\sqrt{N}} + \cdots\right) \\ &= \sqrt{N}\left(0 + \frac{\omega}{\sqrt{N}} + \kappa_3 \frac{\omega^2}{\sqrt{N}} + \cdots\right) \end{aligned}$$

where we used the facts: the first order free cumulant is the mean, and the second order cumulant is variance. As $N \to \infty$, the cumulants that are higher than the second order vanish, thus

$$\lim_{N\to\infty} \sqrt{N}\left(0 + \frac{\omega}{\sqrt{N}} + \kappa_3 \frac{\omega^2}{\sqrt{N}} + \cdots\right) = \omega$$

Using the steps similar to the previous example will find the semicircle distribution. □

5.8.4 Compression of Random Matrix

Consider a $N \times N$ Hermitian matrix $\mathbf{X}$ such that

$$\mathbf{X} = \left[\mathbf{x}_1, \ldots, \mathbf{x}_N\right], \quad \mathbf{x}_i \in \mathbb{C}^N$$

Suppose that the $N \times T$ ($N \leq T$) rectangular matrix $\mathbf{X}_c$ is defined as

$$\mathbf{X}_c = \left[\mathbf{x}_1, \ldots, \mathbf{x}_T\right], \quad \mathbf{x}_i \in \mathbb{C}^N$$

with $c = T/N \leqslant 1$ fixed. It is assumed that we have the R-transform of the $N \times N$ matrix $\mathbf{X}\mathbf{X}^H$ at our disposal. Our problem is to find the R-transform of the $T \times T$ matrix $\mathbf{X}_c^H\mathbf{X}_c$. The idea is to compress the $N \times N$ matrix $\mathbf{X}\mathbf{X}^H$ to the $T \times T$ matrix $\mathbf{X}_c^H\mathbf{X}_c$, by using the project matrix that was treated previously.

Example 5.8.12 (projection matrix for matrix compression) As an example, let the $N \times N$ diagonal matrix $\mathbf{P}$ be a projection matrix such that

$$p_{\mathbf{P}}(x) = (1-c)\,\delta(x) + c\delta(x-1)$$

For $N = 4, c = 1/2$, we have

$$\mathbf{X}\mathbf{X}^H = \begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \\ x_{41} & x_{42} & x_{43} & x_{44} \end{pmatrix}, \quad \mathbf{P}\mathbf{X}\mathbf{X}^H\mathbf{P} = \begin{pmatrix} x_{11} & x_{12} & 0 & 0 \\ x_{21} & x_{22} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

which is called the corner of the matrix $\mathbf{X}\mathbf{X}^H$. It is immediately seen that the eigenvalue distribution of the $N \times N$ corner of the matrix $\mathbf{X}\mathbf{X}^H$ is equivalent to $\mathbf{X}_c^H\mathbf{X}_c$. □

Theorem 5.8.13 (theorem 14.10 in [259]) Consider the $N \times N$ Hermitian random matrix $\mathbf{X}$. Let the $T \times T$ diagonal matrix $\mathbf{P}$ be distributed as

$$p_{\mathbf{P}}(x) = (1-c)\,\delta(x) + c\delta(x-1)$$

Moreover define

$$\mathbf{X}_c = \mathbf{X}\mathbf{P}$$

Then the asymptotic eigenvalue distribution of $\mathbf{X}_c$ converges almost surely to limit

$$p_{\mathbf{X}_c}(x) = p_{\mathbf{X}\mathbf{P}}(x) = (1-c)\,\delta(x) + cp_{\mathbf{Y}}(x)$$

such that the R-transform of $\mathbf{Y}$ satisfies

$$R_{\mathbf{Y}}(\omega) = R_{\mathbf{X}}(c\omega) \tag{5.143}$$

as $N \to \infty$ with $c = T/N \leq 1$ fixed.

Example 5.8.14 (rectangular matrix with i.i.d. entries) Let the entries of the $N \times T$ matrix $\mathbf{X}_c$ be i.i.d with the variance $1/N$ and the ratio $c = T/N \leq 1$ fixed. Then find the R-transform of $\mathbf{X}_c^H\mathbf{X}_c$ for any $c \leq 1$.

With (5.135) for $c = 1$, we have

$$R_{\mathbf{X}_1^H\mathbf{X}_1}(\omega) = \frac{1}{1-\omega}$$

Thus from (5.143), we have

$$R_{\mathbf{X}_c^H \mathbf{X}_c}(\omega) = R_{\mathbf{X}_1^H \mathbf{X}_1}(c\omega) = \frac{1}{1 - c\omega} \qquad \square$$

Theorem 5.8.15 ([271]) Consider an invertible Hermitian matrix $\mathbf{X}$. Then we have

$$\frac{1}{R_{\mathbf{X}}(\omega)} = R_{\mathbf{X}^{-1}}\left(-R_{\mathbf{X}}(\omega)\left(1 + \omega R_{\mathbf{X}}(\omega)\right)\right)$$

5.8.5 Multiplicative Free Convolution

Consider the free random matrices $\mathbf{A}$ and $\mathbf{B}$ and assume that, their asymptotic eigenvalue distributions are known. Now we want to address how to infer the asymptotic eigenvalue distribution of $\mathbf{AB}$.

As defined before, the moment-generating function for a Hermitian random matrix $\mathbf{X}$ is

$$M_{\mathbf{X}}(s) = \sum_{k=1}^{\infty} \phi\left(\mathbf{X}^k\right) s^k \tag{5.144}$$

where $\phi(\cdot)$ is defined in (5.123). Or equivalently

$$M_{\mathbf{X}}(s) = \left(\frac{1}{s}\right) G_{\mathbf{X}}\left(\frac{1}{s}\right) - 1 \tag{5.145}$$

Moreover, the S-transform of $\mathbf{X}$ is

$$S_{\mathbf{X}}(z) = \frac{1+z}{z} M_{\mathbf{X}}^{-1}(s) \tag{5.146}$$

Theorem 5.8.16 (theorem 2.5 in [272]) Let $\mathbf{A}$ and $\mathbf{B}$ be free random matrices such that, either $\phi(\mathbf{A}) \neq 0$ or $\phi(\mathbf{B}) \neq 0$. Then we have

$$S_{\mathbf{AB}}(z) = S_{\mathbf{A}}(z) S_{\mathbf{B}}(z) \tag{5.147}$$

Moreover, R-transform and S-transform has a straightforward relation [259] such that

$$S_{\mathbf{X}}\left(zR_{\mathbf{X}}(z)\right) = \frac{1}{R_{\mathbf{X}}(\omega)} \tag{5.148}$$

Example 5.8.17 (product of two i.i.d. random matrices) Let the entries of the $R \times T$ matrix $\mathbf{X}$ be i.i.d with the variance $1/R$ with the ratio $\beta = T/R$ fixed. Show that

$$S_{\mathbf{X}^H \mathbf{X}}(z) = \frac{1}{1 + \beta z} \tag{5.149}$$

$\square$

Lemma 5.8.18 Consider the $R \times T$ matrix $\mathbf{X}$. Then we have

$$S_{\mathbf{X}\mathbf{X}^H}(z) = \frac{z+1}{z+\beta} S_{\mathbf{X}^H \mathbf{X}}\left(\frac{z}{\beta}\right) \tag{5.150}$$

with $\beta = T/R$.

Example 5.8.19 (*S*-transform of HH^H) Let the entries of the $R \times T$ and $S \times T$ matrices $\mathbf{A}$ and $\mathbf{B}$ be independent identically distributed with zero mean, variances $1/R$ and $1/S$ and

$$\mathbf{H} \triangleq \mathbf{AB} \tag{5.151}$$

Moreover assume that $R, S, T \to \infty$ with ratios $\rho = S/R$ and $\beta = R/T$ fixed. Then find the S-transform of $\mathbf{HH}^H$. Let us first define

$$\mathbf{C}_{R\times R} = \mathbf{ABB}^H\mathbf{A}^H, \qquad \tilde{\boldsymbol{C}}_{S\times S} = \mathbf{A}^H\mathbf{ABB}^H \tag{5.152}$$

where

$$S_{\mathbf{A}^H\mathbf{A}}(z) = \frac{1}{1+\rho z}, \qquad S_{\mathbf{BB}^H}(z) = \frac{1}{z+\beta/\rho}. \tag{5.153}$$

Then the S-transform reads

$$S_{\tilde{C}}(z) = \frac{1}{(1+\rho z)(z+\beta/\rho)}. \tag{5.154}$$

It follows from (5.150) that

$$\begin{aligned} S_{\mathbf{C}}(z) &= \frac{z+1}{z+\rho} S_{\tilde{C}}\left(\frac{z}{\rho}\right) \\ &= \frac{z+1}{z+\rho} \cdot \frac{1}{(1+z)(z+\rho\beta/\rho)} \\ &= \frac{\rho}{(z+\rho)(z+\beta)}. \end{aligned} \tag{5.155}$$

□

Let $\mathbf{X}$ be a $p \times n (p \geq n)$ matrix with standard complex Gaussian entries. The positive definite matrix $\mathbf{X}^H\mathbf{X}$ is then referred to as a complex Wishart matrix. Such matrices are fundamental in random matrix theory. Crucial to these applications is the exact solvability of statistical properties of the eigenvalues of complex Wishart matrices.

Example 5.8.20 (*S*-transform of complex Gaussian and Wishart matrices [75]) For a complex Wishart matrix $\mathbf{X}^H\mathbf{X}$ with $\mathbf{X}$ an $M \geq N$ standard Gaussian matrix we must scale the eigenvalues by dividing $\mathbf{X}^H\mathbf{X}$ by N. With $M \geq N$ fixed, the large N leading eigenvalue support is then the interval $[0, 4]$, and the global density of eigenvalues is given by the Marchenko–Pastur law

$$\rho_{\mathbf{X}^H\mathbf{X}}(x) = \frac{1}{\pi\sqrt{x}}\sqrt{1-x/4}, \quad 0 < x < 4 \tag{5.156}$$

In fact, it is not the global densities themselves that are the central objects of free probability calculus but rather certain transforms.

The most fundamental of these is the Stieltjes transform (a type of Green function)

$$G_{\mathbf{Y}}(z) = \int_I \frac{1}{y-z}\rho_{\mathbf{Y}}(y)dy, \quad z \notin I \tag{5.157}$$

where I denotes the interval of support. From (5.156), we have

$$G_{\mathbf{X}^H\mathbf{X}}(z) = \frac{-1+\sqrt{1-4/z}}{2} \tag{5.158}$$

(see e.g. [62, Exercises 14.4 q.6(i) with $\alpha = 0$]). As the eigenvalues of $\left(\mathbf{X}^H\mathbf{X}\right)^{-1}$ are the reciprocals of the eigenvalues of $\mathbf{X}^H\mathbf{X}$, a straightforward calculation from (5.157) and (5.158) shows

$$G_{(\mathbf{X}^H\mathbf{X})^{-1}}(z) = -\frac{1}{z} - \frac{-1+\sqrt{1-4z}}{2z^2} \tag{5.159}$$

this being a special case of the general relation

$$G_{\mathbf{Y}^{-1}}(z) = -\frac{1}{z} - \frac{G_{\mathbf{Y}}(1/z)}{2z^2} \tag{5.160}$$

Now introduce the auxiliary quantity

$$\Upsilon(z) := -1 - \frac{1}{z}G(1/z) \tag{5.161}$$

so that

$$\Upsilon_{\mathbf{X}^H\mathbf{X}}(z) = -1 - \left(\frac{-1+\sqrt{1-4z}}{2z}\right), \quad \Upsilon_{(\mathbf{X}^H\mathbf{X})^{-1}}(z) = z\left(\frac{-1+\sqrt{1-4/z}}{2}\right)$$

From these explicit forms, we compute the corresponding inverse functions

$$\Upsilon^{-1}_{\mathbf{X}^H\mathbf{X}}(z) = \frac{z}{(1+z)^2}, \quad \Upsilon^{-1}_{(\mathbf{X}^H\mathbf{X})^{-1}}(z) = -\frac{z^2}{1+z} \tag{5.162}$$

Finally, introduce the S-transform by

$$S_{\mathbf{Y}}(z) = \frac{1+z}{z}\Upsilon^{(-1)}_{\mathbf{Y}}(z) \tag{5.163}$$

We see from (5.162) that

$$S_{\mathbf{X}^H\mathbf{X}}(z) = \frac{1}{1+z}, \quad S_{(\mathbf{X}^H\mathbf{X})^{-1}}(z) = -z \tag{5.164}$$

□

Theorem 5.8.21 (the law of large numbers for the free additive convolution of measures with bounded support [273]) Let μ be a probability measure on $\mathbb{R}$ with existing mean value α, and let $\psi_n : \mathbb{R} \to \mathbb{R}$ be the map $\psi_n(x) = \frac{1}{n}x$. Then

$$\frac{d\psi_n(x)}{dx}(\mu \boxplus \cdots \boxplus \mu) \to \delta_\alpha$$

where convergence is weak and δ_x denotes the Dirac measure at $x \in \mathbb{R}$.

Here $\frac{d\phi(\mu)}{dx}$ denotes the image measure of μ under ϕ for a Borel measurable function $\phi : \mathbb{R} \to \mathbb{R}$, respectively, $[0,\infty) \to [0,\infty)$.

Theorem 5.8.22 (The free multiplicative law for measures with unbounded support [274]) Let μ be a probability measure on $[0,\infty)$ and let $\phi_n : [0,\infty) \to [0,\infty)$ be the map $\phi_n(x) = x^{1/n}$. Set $\delta = \mu(\{0\})$. If we denote

$$\nu_n = \frac{d\phi_n(\mu_n)}{dx} = \frac{d\phi_n}{dx}\left(\overbrace{\mu \boxtimes \cdots \boxtimes \mu}^{n \text{ times}}\right)$$

then ν_n converges weakly to a probability measure ν on $[0,\infty)$. If μ is a Dirac measure on $[0,\infty)$, then $\nu=\mu$. Otherwise ν is the unique measure on $[0,\infty)$ characterized by $\nu\left(\left[0,\frac{1}{S_\mu(t-1)}\right]\right)=t$ for all $t\in(\delta,1)$ and $\nu(\{0\})=\delta$. The support for the measure ν is the closure of the interval

$$(a,b)=\left(\left(\int_0^\infty x^{-1}d\mu(x)\right)^{-1},\int_0^\infty xd\mu(x)\right)$$

where $0\leqslant a\leqslant b\leqslant\infty$.

Note that, unlike in the additive case, the multiplicative limit distribution is only a Dirac measure if μ is a Dirac measure. Furthermore S_μ and hence ([269, Theorem 2.6]) μ can be reconstructed from the limit measure.

Proposition 5.8.23 (two-parameter family of measures [274]) Let $\alpha,\beta\geq 0$. There exists a probability measure $\mu_{\alpha,\beta}$ on $(0,\infty)$ for which the S-transform is given by

$$S_{\mu_{\alpha,\beta}}(z)=\frac{(-z)^\beta}{(1+z)^\alpha},\quad 0<z<1,\alpha>0,\beta>0 \tag{5.165}$$

Furthermore, these measures form a two-parameter semigroup, multiplicative under $\boxtimes$ induced by multiplication of $(\alpha,\beta)\in[0,\infty)\times[0,\infty)$.

Example 5.8.24 (two-parameter family of measures [274]) For $(\alpha,\beta)=(1,0)$, we have $S_{\mu_{1,0}}(z)=\frac{1}{1+z}$, which the S-transform of the free Poisson distributions with shape parameter 1 (also called the Marchenko–Pastur law). The distribution is give by

$$\mu_{1,0}(x)=\frac{1}{2\pi}\sqrt{\frac{4-x}{x}}\mathbb{I}_{(0,4)}(x)\,dx$$

where $\mathbb{I}$ is the indicator function. □

We can use (5.165) to model the massive datasets, where two parameters $\alpha,\beta\geq 0$ must be estimated from the data.

5.9 Random Vandermonde Matrix

A Vandermonde matrix with entries on the unit circle has the following form:

$$\mathbf{V}=\frac{1}{\sqrt{N}}\begin{pmatrix}1 & \cdots & 1\\ e^{-j\omega_1} & \cdots & e^{-j\omega_L}\\ \vdots & \ddots & \vdots\\ e^{-j(N-1)\omega_1} & & e^{-j(N-1)\omega_L}\end{pmatrix}_{N\times L} \tag{5.166}$$

We will consider the case where $\omega_1,\ldots,\omega_L$ are independent and identically distributed (i.i.d.), taking values in $[0,2\pi]$. Throughout this section, the ω_i will be called phase distributions. $\mathbf{V}$ will be used only to denote Vandermonde matrices with a given phase distribution, and the dimensions of the Vandermonde matrices will always be $N\times L$.

In many practical applications, N and L are quite large, and we may be interested in studying the case where both go to a given ratio $\frac{L}{N} \to c$. The factor $\frac{1}{\sqrt{N}}$, as well as the assumption that the Vandermonde entries $e^{-j\omega_i}$ lie on the unit circle, are included in (5.166) to ensure that the analysis will give limiting asymptotic behavior. In general, often the moments, not the moments of the determinants, are the quantities we seek. It can be shown that, asymptotically, the moments of the Vandermonde matrices depend only on the ratio c and the phase distribution, and have explicit expressions. Moments are useful for performing deconvolution.

The fact that all the moments exist is not enough to guarantee that there exists a limit probability measure having these moments. However, we will prove that, in this case, this is true. In other words, the matrices $\mathbf{V}^H\mathbf{V}$ converge in distribution to a probability measure μ_c supported on $[0, +\infty]$ where $c = \lim \frac{L}{N}$ as the dimension grows. More precisely, let μ_L be the empirical eigenvalue distribution of the random matrices $\mathbf{V}^H\mathbf{V}$. Then μ_L converge weakly to a unique probability measure μ_c supported in $[0, +\infty]$ with moments

$$m_n^{(c)} = \int_0^{+\infty} t^n d\mu_c(t)$$

We also enlarge the class of functions for which the limit eigenvalue distribution exists to include unbounded densities and we find lower bounds and upper bounds for the maximum eigenvalue.

Example 5.9.1 (capacity of the Vandermonde channel [275]) Consider the Gaussian matrix channel in which the received signal $\mathbf{y} \in \mathbb{C}^{N\times 1}$ is given as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{z} \tag{5.167}$$

where $\mathbf{z} \sim \mathcal{CN}\left(\mathbf{0}, \mathbf{I}_N\right)$, $\mathbf{x} \sim \mathcal{CN}\left(\mathbf{0}, \mathbf{I}_L\right)$, $\mathbf{H} \in \mathbb{C}^{N\times L}$ has i.i.d. zero mean Gaussian entries and is standard. Then an explicit expression for the asymptotic capacity exists

$$\begin{aligned}&\lim_{N\to\infty} \log\det\left(\mathbf{I}_N + \gamma\mathbf{H}\mathbf{H}^H\right)\\ &\quad = -\frac{\log e}{4\gamma}F(\gamma,\beta) + \beta\log\left(1+\gamma-\frac{1}{4}F(\gamma,\beta)\right) + \log\left(1+\beta\gamma-\frac{1}{4}F(\gamma,\beta)\right)\end{aligned}$$

where

$$F(a,b) := \left(\sqrt{a\left(1+\sqrt{b}\right)^2+1} - \sqrt{a\left(1-\sqrt{b}\right)^2+1}\right)^2$$

and the SNR γ is

$$\gamma = \frac{N\cdot\mathbb{E}\left[\|\mathbf{x}\|^2\right]}{L\cdot\mathbb{E}\left[\|\mathbf{z}\|^2\right]}$$

and the ratio $\frac{N}{L} \to \beta$ as $N \to \infty$.

We can prove that a similar limit exists and is finite if the Gaussian matrix is replaced with a random Vandermonde matrix. Besides, using Jensen's inequality, we can get an upper bound on the capacity. More precisely, if we fix the SNR Y, we may define the

asymptotic capacity of the Vandermonde channel (whenever the limiting moments exist and define a measure) for random Vandermonde matrices $\mathbf{V} \in \mathbb{C}^{N\times L}$, to be

$$\begin{aligned} C_{\mathbf{V}}(\gamma) &:= \lim_{N\to\infty} \mathbb{E}\left(\frac{1}{N}\log\det\left(\mathbf{I}_N + \gamma\mathbf{V}\mathbf{V}^H\right)\right) \\ &= \lim_{N\to\infty} \mathbb{E}\left(\frac{1}{N}\log\det\left(\mathbf{I}_L + \gamma\mathbf{V}^H\mathbf{V}\right)\right) \\ &= \lim_{N\to\infty} \mathbb{E}\left(\frac{1}{N}\mathrm{Tr}\log\left(\mathbf{I}_L + \gamma\mathbf{V}^H\mathbf{V}\right)\right) \\ &= \lim_{L\to\infty} \int_0^\infty c\log(1+\gamma t)\, d\mu_L(t) \\ &= \int_0^\infty c\log(1+\gamma t)\, d\mu_c(t) \end{aligned} \tag{5.168}$$

where μ_L is the empirical measure of the $L \times L$ random matrix $\mathbf{V}^H\mathbf{V}$ and μ_c is the limit measure of the μ_L. The first equality follows from Sylvester's theorem on determinants, the second and third are by definition, and the final equality is a consequence of their uniform integrability. This latter follows from $\log(1+\gamma t) < \gamma t, t > 0$ and that given $\varepsilon > 0, \exists\alpha > 0$ such that

$$\sup_L \int_\alpha^\infty t d\mu_L(t) < \varepsilon$$

see the converse statement in [276, Theorem 5.4]. Therefore, by Jensen's inequality

$$\begin{aligned} C_{\mathbf{V}}(\gamma) &= \int_0^\infty c\log(1+\gamma t)\, d\mu_c(t) \\ &\leqslant c\log(1+\gamma) \end{aligned}$$

since the limit first moment is 1.

Consider an application for a network with M mobile users conducting synchronous multiaccess to a base station with N antenna elements, arranged as a uniform linear

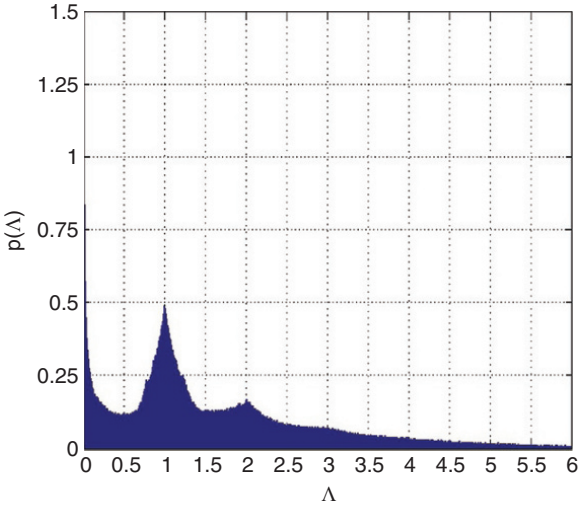


Figure 5.2 Simulated limit distribution for a uniform distribution $\theta \sim U[-\pi, \pi]$ with $L = N = 1000$ averaged over 700 sample matrices. Source: Reproduced from [275] with permission.

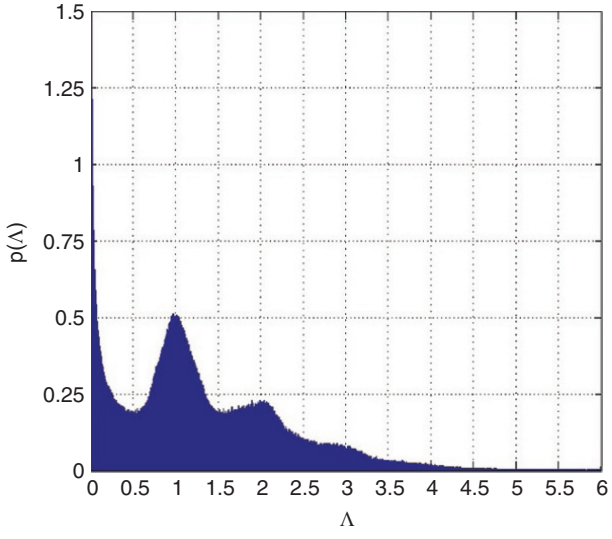


Figure 5.3 Simulated limit distribution for $\theta \sim f(x)$, where $f(x)$is an unbounded pdf defined as $f(x) = \frac{1}{2\pi} \log \frac{\pi}{|x|}$. The distributions with $L = N = 1000$ are averaged over 700 sample matrices. Source: Reproduced from [275] with permission.

array. Suppose a random subset of L users in any time slot are selected to transmit. Then the antenna array response over the selected users is $\mathbf{V}$ defined as

$$\mathbf{V} = \frac{1}{\sqrt{N}} \begin{pmatrix} 1 & \cdots & 1 \\ e^{-i2\pi d/\lambda \sin(\theta_1)} & \cdots & e^{-i2\pi d/\lambda \sin(\theta_L)} \\ \vdots & \ddots & \vdots \\ e^{-i2\pi(N-1)d/\lambda \sin(\theta_1)} & & e^{-i2\pi(N-1)d/\lambda \sin(\theta_L)} \end{pmatrix}_{N \times L} \tag{5.169}$$

where d is the element spacing and λ is the wavelength. Let us assume that M, L, N are large and that the angles of arrival are uniformly scattered in $(-\alpha, \alpha)$. Then it is reasonable to determine the performance with the assumption that the angles of arrival are drawn *uniformly* so that the maximum sum throughput (equivalently per user rate) is determined by (5.168) with the phase pdf given as

$$q_\alpha(\theta) = \frac{1}{2\beta \sqrt{\frac{4\pi^2 d^2}{\lambda^2} - \theta^2}}$$

for $\theta \in \left[-\frac{2\pi d}{\lambda} \sin(\alpha), \frac{2\pi d}{\lambda} sin(\alpha) \right]$.

An unbounded pdf is defined as

$$f(x) = \frac{1}{2\pi} \log \frac{\pi}{|x|} \tag{5.170}$$

For an illustration of Eq. (5.170), see Figure 5.2 and Figure 5.3. □

Example 5.9.2 (massive MIMO) We study the spatial degrees of freedom of multiple-input multiple-output (MIMO) transmissions in a wireless network with

homogeneously distributed nodes, under the following classical line-of-sight propagation model between node k and node j in the network [277, 278]:

$$h_{jk} = \frac{\exp\left(i2\pi r_{jk}/\lambda\right)}{r_{jk}} \tag{5.171}$$

In the above equation, λ is the carrier wavelength and r_{jk} is the internode distance. From a mathematical point of view, these matrices are interesting objects, as they are halfway between purely random matrices with i.i.d. entries and fully deterministic matrices. Indeed, the internode distances r_{jk} are random due to the random node positions, but there is a clear correlation between the matrix entries. □

5.10 Non-Asymptotic Analysis of State Estimation

State estimation may be formulated as random vector channels. We show that our nonasymptotic framework can be applied in Bayesian inference problems that involve estimation of uncorrelated components. Consider a simple *linear* statistical model

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{z} \tag{5.172}$$

where $\mathbf{H} \in \mathbb{C}^{n\times p}$ is a known matrix, $\mathbf{x} \in \mathbb{C}^p$ denotes an input signal with zero mean and covariance matrix $P\mathbf{I}_p$ and $\mathbf{z}$ represents a zero-mean noise uncorrelated with $\mathbf{x}$, which has covariance $\sigma^2\mathbf{I}_n$. For an arbitrary random vector $\mathbf{x}'$ with a covariance matrix $\mathbf{\Sigma}_x$, we can always normalize such that $\mathbf{x} = \left(\mathbf{\Sigma}_x\right)^{-1/2}\mathbf{x}'$ so the resultant covariance matrix of $\mathbf{x}$ is $P\mathbf{I}_p$. In this section, we assume that

$$\alpha := \frac{p}{n} \leqslant 1 \tag{5.173}$$

i.e. the sample size exceeds the dimension of the input signal. The MMSE estimate of $\mathbf{x}$ given $\mathbf{y}$ can be expressed as [279]

$$\begin{aligned}\hat{\boldsymbol{x}} &= \mathbb{E}\left(\mathbf{x}|\mathbf{y}\right) = \mathbb{E}\left(\mathbf{x}\mathbf{y}^H\right)\left(\mathbb{E}\left(\mathbf{x}\mathbf{y}^H\right)\right)^{-1}\mathbf{y} \\ &= P\mathbf{H}^H\left(\sigma^2\mathbf{I}_n + P\mathbf{H}\mathbf{H}^H\right)^{-1}\mathbf{y}\end{aligned} \tag{5.174}$$

and the resulting MMSE is given by

$$\mathrm{MMSE}\left(\mathbf{H}\right) = \mathrm{Tr}\left(P\mathbf{I}_p - P^2\mathbf{H}^H\left(\sigma^2\mathbf{I}_n + P\mathbf{H}\mathbf{H}^H\right)^{-1}\mathbf{H}\right) \tag{5.175}$$

As a result, the normalized MMSE (NMMSE) can be written as

$$\mathrm{NMMSE}\left(\mathbf{H}\right) := \frac{MMSE\left(\mathbf{H}\right)}{\mathbb{E}\|\mathbf{x}\|^2} = \mathrm{Tr}\left(\mathbf{I}_p - \mathbf{H}^H\left(\frac{1}{\mathrm{SNR}}\mathbf{I}_n + \mathbf{H}\mathbf{H}^H\right)^{-1}\mathbf{H}\right) \tag{5.176}$$

where $\mathrm{SNR} := \frac{P}{\sigma^2}$. We can evaluate the function $\mathrm{NMMSE}\left(\mathbf{H}\right)$) in a reasonably tight manner, as stated below. We need to define a scalar-valued function

$$f\left(\delta, \mathbf{H}\right) := \frac{1}{p}\mathbb{E}\,\mathrm{Tr}\left(\left(\delta + \mathbf{H}\mathbf{H}^H\right)^{-1}\right)$$

Suppose that $\mathbf{H} = \mathbf{AM}$, where $\mathbf{A} \in \mathbb{C}^{n\times m}$ is a deterministic matrix for some integer $m \geq p$, and $\mathbf{M} \in \mathbb{C}^{n\times p}$ is a random matrix such that M_{ij}s are independent random variables satisfying $\mathbb{E}M_{ij} = 0$ and $\mathbb{E}\left|M_{ij}\right|^2 = \frac{1}{p}$.

If $\sqrt{p}M_{ij}$s are bounded by D, then for any $t > 8\sqrt{\frac{\pi}{p}}$, one has

$$\frac{1}{p}\,\mathrm{NMMSE}\,(\mathbf{H}) \in \left[f\left(\frac{8}{9\,\mathrm{SNR}}, \mathbf{H}\right) + \tau_{\mathrm{bd}}^{\mathrm{lb}}, f\left(\frac{9}{8\,\mathrm{SNR}}, \mathbf{H}\right) + \tau_{\mathrm{bd}}^{\mathrm{ub}}\right] \tag{5.177}$$

with probability exceeding $1 - 8\exp\left(-\frac{pt^2}{16}\right)$. Here

$$\tau_{\mathrm{bd}}^{\mathrm{lb}} = -\frac{2\sqrt{2}tD\,\|\mathbf{A}\|\,\mathrm{SNR}^{1.5}}{3\sqrt{3}p}, \qquad \tau_{\mathrm{bd}}^{\mathrm{ub}} = \frac{3\sqrt{3}tD\,\|\mathbf{A}\|\,\mathrm{SNR}^{1.5}}{8p}$$

We can obtain similar results if H_{ij}s satisfy the logarithmic Sobolev inequality. This is also true when H_{ij}s are independently drawn from heavy-tailed distributions.

Bibliographical Remarks

We have drawn on material from [1] at the beginning of this chapter. In several sections, we use materials from [82, 147].

We follow [121] in the exposition of Section 5.2.1.

Section 5.3 is taken from the review paper [280] .

Section 5.6 and Section 5.7 are taken from the review paper [176].

In this whole chapter, we have borrowed material from [139].

In Section 5.9, the key reference on the random Vandermonde matrix is [281]. A good overview of analytical tools for large random matrices in communications [282]. Tucci and Whiting (2011, 2012) [275,283] treat the topic in mathematical depth. Recently this topic was studied in the context of massive MIMO [277, 278, 284, 285]. The 5G wireless network will have impact on smart grid, the next generation power grid. We follow [281] for the exposition of the basic properties.

A nonasymptotic analysis of random vector channels has been performed in [40]. In particular, this is done in a systematic manner for the detection of extremely weak signals. The foundation of the nonasymptotic analysis is the concentration of spectral measure phenomenon. In Section 5.10, we give some complementary treatment of this topic, following [286] for the exposition.

6

Large Non-Hermitian Random Matrices and Quatartenionic Free Probability Theory

This chapter studies (large) non-Hermitian random matrices using the newly developed quatartenionic free probability theory. In our opinion, this development will be the new paradigm to represent large datasets and new big data analytics will be derived with it. Most results are appearing here in book form for the first time. Some results have never appeared in a publication before. This chapter is also the culmination of the random matrix theory development. The most important fact is that non-Hermitian random matrices have complex-valued eigenvalues. As shown in Section 1.4, the new concept of free entropy—defined in the complex plane—is introduced to define the "information."

Recently, for example in [287], products of random matrices have experienced a revival due to new mathematical insights about the statistics of the eigen- and singular values for finite as well as infinite matrix dimensions. Due to recent progress in the field, now we may study a product of an arbitrary number of random matrices of arbitrary size for certain matrix ensembles. Since the number of matrices and their size can be chosen freely, discussions of various limits are allowed. This not only includes macroscopic (as well as microscopic) structures for infinite matrix dimension, but also available is the limit where the number of matrices goes to infinite. Analogous to the study of individual random matrices, products of random matrices show a *rich mathematical structure* and various limits have revealed new universality classes, which are important in the physical sciences as well as in mathematics and beyond.

Our interest in products of large random matrices arises from the rich mathematical structure of this objects. We believe that knowledge discovery is to find the structure behind the large datasets. This novel mathematical object is very natural in the context of big data. Often, we are interested in a matrix-valued time series that are conveniently modeled as a sequence of $N \times N$ matrices $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_L$. We are interested in the large random matrices, say $N = 100 - 1000$. These matrix-valued random variables are building blocks for the big data problem at hand. Basic matrix operations include:

- adding up the L matrices $\mathbf{A}_L = \mathbf{X}_1 + \mathbf{X} + \cdots + \mathbf{X}_L$;
- products of L matrices $\mathbf{P}_L = \mathbf{X}_1\mathbf{X}_2 \cdots \mathbf{X}_L$;
- geometric mean of L matrices $\left(\mathbf{P}_L\right)^{1/L} = \left(\mathbf{X}_1\mathbf{X}_2 \cdots \mathbf{X}_L\right)^{1/L}$;
- $\mathbf{X}_1^{1/M}\mathbf{X}_2^{1/M} \cdots \mathbf{X}_L^{1/M}$ for non-negative integer $M \geq 1$.

The most useful observation is to preserve the symmetry of eigenvalues on the complex plane. For example, for a complex number in its polar form $z = |z|\, e^{j\phi}$, the function

$$z^{\alpha} \equiv \exp\left(\alpha \log(z)\right) = |z|^{\alpha} e^{j\alpha\phi}, \quad \text{for } \alpha \in \mathbb{R}$$

Smart Grid using Big Data Analytics: A Random Matrix Theory Approach, First Edition.
Robert C. Qiu and Paul Antonik.

has the property of preserving the symmetry. In particular, $\alpha = L/M$ for non-negative integers L and M.

For large random matrices, it is very interesting to discover that the product $\mathbf{X}_1\mathbf{X}_2\cdots\mathbf{X}_L$ behaves as the power $\mathbf{X}^L$ of a single matrix $\mathbf{X}$, for $\mathbf{X} = \mathbf{X}_i, i = 1, \ldots, L$.

Outliers to the circular law, the (single) ring law, and the elliptic law are studied. In practice, the ring law is most important since the rectangular random matrix is naturally encountered. The circular law is the special case of the single ring law. Associated with the single ring law is the random singular value decomposition (SVD). MATLAB codes are included to gain hands-on insight.

6.1 Quatartenionic Free Probability Theory

Non-Hermitian matrices have complex-valued eigenvalue distribution in general. In the Hermitian case, we work on the complex-valued matrix functions to search real-valued eigenvalues, while we now have to work on a q-valued function to search complex-valued eigenvalues (see Figure 6.1), where q is defined in (6.10). See also See Table 6.1.

Since non-Hermitian matrices have real eigenvalues, the method of analysis to deal with real-valued eigenvalue distributions in free probability is to use complex analysis, that is to represent a real-valued eigenvalue distribution

$$p(x) = \frac{1}{\pi}\lim_{\varepsilon\to 0+}\mathrm{Re}\left\{jG\left(x+j\varepsilon\right)\right\} \tag{6.1}$$

as a limit of a complex-valued holomorphic function $G(s)$, which is the Stieltjes transform defined by

$$G(s) = \int \frac{1}{s-x}dP(x) \tag{6.2}$$

where $p(x) = dP(x)/dx$.

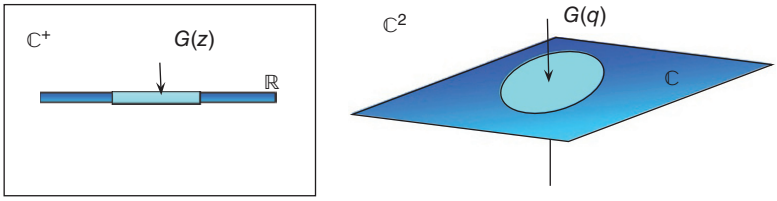


Figure 6.1 Left: Complex-valued operation for a real function in upper complex plane. Right: Quaternion-valued operation for a complex function in hyper complex plane.

Table 6.1 Comparison between classical, free, and quatartenionic free probability theories.

	Probability space	Algebra
Classical probability	Commutative	Commutative
Free probability	Non-commutative	Commutative
Quatartenionic free probability	Non-commutative	Non-commutative

Complex-valued eigenvalue distributions are often circularly symmetric and, thus, not holomorphic. They can be represented by a pair of holomorphic functions representing real and imaginary part. Instead of real and imaginary part of a complex variable z, one can also consider z, its complex conjugate z^*, and apply the Wirtinger rule [288] for differentiation:

$$\frac{\partial z}{\partial z^*} = 0 = \frac{\partial z^*}{\partial z} \tag{6.3}$$

6.1.1 Stieltjes Transform

In order to generalize the Stieltjes transform to two complex variables z and z^*, we first rewrite (6.2) by

$$G(s) = \frac{d}{ds}\int \log(s-x)dP(x) \tag{6.4}$$

We further note that the Dirac function of a complex argument can be represented as the limit

$$\begin{aligned}\delta(z-z') &= \tfrac{1}{\pi}\lim_{\varepsilon\to 0}\tfrac{\varepsilon^2}{\left(|z-z'|^2+\varepsilon^2\right)^2}\\ &= \tfrac{1}{\pi}\lim_{\varepsilon\to 0}\tfrac{\partial^2}{\partial z\partial z^*}\log\left[|z-z'|^2+\varepsilon^2\right]\end{aligned} \tag{6.5}$$

Thus we obtain

$$p(z) = \frac{1}{\pi}\lim_{\varepsilon\to 0}\frac{\partial^2}{\partial z\partial z^*}\int \log\left[\left|z-z'\right|^2+\varepsilon^2\right]dP(z) \tag{6.6}$$

We define the bivariate Stieltjes transform by

$$\begin{aligned}G(s,\varepsilon) &= \frac{\partial}{\partial s}\int \log\left[|s-z|^2+\varepsilon^2\right]dP(z)\\ &= \int \frac{(s-z)^*}{|s-z|^2+\varepsilon^2}dP(z)\end{aligned} \tag{6.7}$$

and get the bivariate Stieltjes inversion formula to read

$$p(z) = \frac{1}{\pi}\lim_{\varepsilon\to 0}\frac{\partial}{\partial z^*}G(s,\varepsilon). \tag{6.8}$$

At first sight, the bivariate Stieltjes transform looks quite different from (6.2). We can, however, rewrite (6.7) as

$$G(s,\varepsilon) = \int\left[\begin{pmatrix} s-z & i\varepsilon \\ i\varepsilon & s^*-z^*\end{pmatrix}^{-1}\right]_{11} dP(z) \tag{6.9}$$

which clearly resembles the form of (6.2). To get an even more striking analogy with (6.2), we can introduce the Stieltjes transform with quaternionic argument $q \equiv \nu + j\omega, \quad (\nu,\omega)\in\mathbb{C}^2, \quad i^2\equiv -1, \quad ij = ji$

$$G(q) \equiv \int \frac{1}{q-z}dP(z) \tag{6.10}$$

and with the respective inversion formula

$$p(z) = \frac{1}{\pi}\lim_{\varepsilon\to 0}\frac{\partial}{\partial z^*}\Re G(z+i\varepsilon) \tag{6.11}$$

and the definition $\mathfrak{R}\left(v+j\omega\right) \equiv v \in \mathbb{C}^2$. Note that real and imaginary part of a quaternion are its first and second complex components, respectively. Quaternions are inconvenient to deal with, since multiplication of quaternions does not commute, in general. However, any quaternion $q = v + j\omega$ can be conveniently represented by the complex-valued 2×2 matrix

$$\begin{pmatrix} v & \omega \\ -\omega^* & v^* \end{pmatrix} \tag{6.12}$$

This matrix representation directly connects (6.9) with (6.10) via

$$G(s,\varepsilon) = \mathfrak{R}G(s+i\varepsilon). \tag{6.13}$$

Finally, the quaternion-valued Stieltjes transform can be expressed as

$$\begin{aligned} G(q) &\equiv \int \left(1-q^{-1}z\right)^{-1} q^{-1} dP(z) \\ &= \sum_{k=0}^{\infty} \int \left(q^{-1}z\right)^k q^{-1} dP(z) \\ &= \sum_{k=0}^{\infty} \mathbb{E}\left[\left(q^{-1}z\right)^k\right] q^{-1} \end{aligned} \tag{6.14}$$

Note that the last equation of (6.14) is equivalent to

$$q^{-1} \sum_{k=0}^{\infty} \mathbb{E}\left[\left(q^{-1}z\right)^k\right] \tag{6.15}$$

We follow (6.14) for the rest of the work.

6.1.2 Additive Free Convolution

We define the S-transform of quaternion argument p in complete analogy to the complex case in [52] as

$$R(p) = G^{-1}(p) - p^{-1} \tag{6.16}$$

and obtain for free random matrices $\mathbf{A}$ and $\mathbf{B}$, with $R_{\mathbf{A}}(p)$ and $R_{\mathbf{B}}(p)$ denoting the R-transforms of the respective asymptotic eigenvalue distributions,

$$R_{\mathbf{A}+\mathbf{B}}(p) = R_{\mathbf{A}}(p) + R_{\mathbf{B}}(p) \tag{6.17}$$

The scaling law of the R-transform is generalized as

$$R_{z\mathbf{A}}(p) = zR_{\mathbf{A}}(pz) \tag{6.18}$$

for $z \in \mathbb{C}$. Note that the order of factors does matter here, since $pz \neq zp$, in general.

Let $\mathbf{A}$ and $\mathbf{B}$ are free each others. Then we have

$$G_{\mathbf{A}+\mathbf{B}}(q) = G_{\mathbf{A}}\left(q - R_{\mathbf{B}}\left[G_{\mathbf{A}+\mathbf{B}}(q)\right]\right) \tag{6.19}$$

(6.19) can be derived as follows.

$$\begin{aligned} q &= G_{\mathbf{A}}\left[G_{\mathbf{A}}^{-1}(q)\right] \\ &= G_{\mathbf{A}}\left[G_{\mathbf{A}+\mathbf{B}}^{-1}(q) - G_{\mathbf{B}}^{-1}(q) + \tfrac{1}{q}\right] \\ &= G_{\mathbf{A}}\left[G_{\mathbf{A}+\mathbf{B}}^{-1}(q) - R_{\mathbf{B}}(q)\right]. \end{aligned} \tag{6.20}$$

By substitution $q \to G_{\mathbf{A}+\mathbf{B}}(q)$, we have

$$G_{\mathbf{A}+\mathbf{B}}(q) = G_{\mathbf{A}}\left(q - R_{\mathbf{B}}\left[G_{\mathbf{A}+\mathbf{B}}(q)\right]\right).$$

6.1.3 Multiplicative Free Convolution

The key quantity of interest in random matrix theory is the eigenvalue density, which may be equivalently expressed through the Green's function. The R and S transforms satisfy functional relations with the Green's function and hence their knowledge is equivalent (in the Hermitian case) to the knowledge of the eigenvalue density (or more precisely of its moments).

While additive free convolution generalizes straightforwardly, this is very different for multiplicative free convolution.

The Green's function for non-Hermitian matrices is conveniently expressed as 2×2 matrices with complex elements. In order to distinguish this situation from the Hermitian case, where functions and their arguments were complex numbers, we shall use calligraphic letters to denote the corresponding 2×2 complex matrices. The R transform in this case is a map of a space of 2×2 complex matrices onto a space of 2×2 complex matrices $\mathcal{G} \to R(\mathcal{G})$.

We define a modified quaternion-valued Stieltjes transform of a non-Hermitian random matrix $\mathbf{X}$ as

$$\mathcal{G}_{\mathbf{X}} = \lim_{\varepsilon \to 0} G_{\mathbf{X}}(z + i\varepsilon) \tag{6.21}$$

Moreover, for any $q \in \mathbb{C}^2$, we define the following operation as

$$q^L = \omega q \omega^*, \qquad q^R = \omega^* q \omega \tag{6.22}$$

where $\omega \triangleq e^{(j \arg z)/4}$. Let the non-Hermitian matrices $\mathbf{A}$ and $\mathbf{B}$ be free from each other. Then we have [289]

$$R_{\mathbf{AB}}\left(\mathcal{G}_{\mathbf{AB}}\right) = \left[R_{\mathbf{A}}\left(\mathcal{G}_{\mathbf{B}}\right)\right]^L \cdot \left[R_{\mathbf{B}}\left(\mathcal{G}_{\mathbf{A}}\right)\right]^R \tag{6.23}$$

But this is a nontrivial formula and very less fruitful, compareing with quaternion valued R transform. On the other hand, there is an interesting result in the name of S-transform over (noncommutative) unital Banach algebra [290].

6.1.4 Quaternion-valued Functions for Hermitian Matrices

Recall that the quaternion-valued Stieltjes can be expanded as

$$G(q) = \int \frac{1}{q - z} dP(z) = \sum_{k=0}^{\infty} \mathbb{E}\left[\left(q^{-1} z\right)^k\right] q^{-1}$$

The quaternion-valued Stieltjes transform for a real distribution, however, can be written as

$$\begin{aligned} \int \tfrac{1}{q-x} dP(x) &= \sum_{k=0}^{\infty} \int \tfrac{x^k}{q^{k+1}} dP(z) \\ &= \sum_{k=0}^{\infty} \tfrac{m_k}{q^{k+1}} \end{aligned} \tag{6.24}$$

since $qx = xq, x \in \mathbb{R}$. This yields the same algebra as in the complex case. Therefore, quaternion-valued Stieltjes, R and S transforms for a real distribution are simply equivalent to the complex case. Obviously

$$G_{\mathbf{H}}(q) = G_{\mathbf{H}}(s)\Big|_{s=q}, \qquad R_{\mathbf{H}}(p) = R_{\mathbf{H}}(\omega)\Big|_{\omega=q}, \qquad S_{\mathbf{H}}(r) = S_{\mathbf{H}}(z)\Big|_{z=r} \tag{6.25}$$

where $\mathbf{H}$ is a Hermitian matrix.

Example 6.1.1 (semicircle and full circle elements) Let $\mathbf{H}$ is a semicircle element. Find $G_{\mathbf{H}}(q)$ and $R_{\mathbf{H}}(q)$. Since odd moments of an even distribution vanish let C_k be the k-th Catalan number, We have

$$\int \frac{1}{q-x} dP_{\mathbf{H}}(x) = \sum_{k=0}^{\infty} \frac{1}{q^{2k+1}} C_k \tag{6.26}$$

By using the recursive expression of the Catalan number, we have [259]

$$\begin{aligned} G_{\mathbf{H}}(q) &= \frac{1}{q} + \sum_{k=1}^{\infty} \frac{1}{q^{2k+1}} \left(\sum_{m=1}^{k} C_{m-1} C_{k-m} \right) \\ &= q^{-1} + q^{-1} \sum_{k=1}^{\infty} \sum_{m=1}^{k} \frac{C_{m-1}}{q^{2k+1}} \cdot \frac{C_{k-m}}{q^{2(m-k)+1}} \\ &= q^{-1} + q^{-1} \sum_{m=1}^{\infty} \frac{C_{m-1}}{q^{2m+1}} \cdot \left(\sum_{k=m}^{\infty} \frac{C_{k-m}}{q^{2(m-k)+1}} \right) \\ &= q^{-1} + q^{-1} \sum_{m=1}^{\infty} \frac{C_{m-1}}{q^{2m+1}} \cdot G_{\mathbf{H}}(q) \\ &= q^{-1} + q^{-1} G_{\mathbf{H}}^2(q) = q^{-1} \left(1 + G_{\mathbf{H}}^2(q)\right) \end{aligned} \tag{6.27}$$

which leads to the following solution:

$$G_{\mathbf{H}}^2(q) = \frac{1}{2} \left[q - \left(q^2 - 4\right)^{1/2} \right]$$

Now substituting q for $G_{\mathbf{H}}^{-1}(q)$ in the last identity of (6.27), we have

$$\begin{aligned} 0 &= G_{\mathbf{H}}^{-1}(q) \left(1 + q^2\right) - q \\ &= \left[R_{\mathbf{H}}(q) + q^{-1} \right]^{-1} \left(1 + q^2\right) - q \\ &= \left(1 + q^2\right) - \left[R_{\mathbf{H}}(q) + q^{-1} \right]^{-1} q \end{aligned}$$

which gives

$$R_{\mathbf{H}}(q) = q \tag{6.28}$$

Let $\mathbf{G}$ is a full circle element, then we have

$$R_{\mathbf{G}}(q) = \Im q \tag{6.29}$$

(6.29) can be derived as follows. $\mathbf{G}$ can be decomposed as

$$\mathbf{G} = \frac{\mathbf{H}_1 + \mathbf{H}_2}{\sqrt{2}}$$

where $\mathbf{H}_1$ and $\mathbf{H}_2$ are two semicircle elements and free from each other. Then we have

$$R_{\mathbf{G}}(q) = \frac{1}{2} \left(q + jqj \right) = \Im q$$

where $\Im q = \Im \left(v + j\omega \right) \equiv w \in \mathbb{C}^2$. □

6.2 R-diagonal Matrices

Definition 6.2.1 (Biane and Lehner (2001) [291]) A random matrix $\mathbf{X}$ is called R-diagonal if it can be decomposed as $\mathbf{X} = \mathbf{UY}$, such that $\mathbf{U}$ is Haar unitary and free of $\mathbf{Y} = \sqrt{\mathbf{X}\mathbf{X}^H}$.

As the matrix size grows, independence is converted into freeness according to some freeness result; see Hiai and Petz (2006) [169]. Therefore, bi-unitarily invariant matrices are asymptotically R-diagonal. Note that independent R-diagonal matrices are free of each other.

R-diagonal matrices have circularly symmetric eigenvalue distribution. In order to determine the boundary of such distributions, we define the following measures [292]:

$$\begin{aligned} in(\mathbf{X})^2 &\triangleq \int \frac{1}{x} dF_{\mathbf{X}\mathbf{X}^H}(x) \\ out(\mathbf{X})^2 &\triangleq \int x dF_{\mathbf{X}\mathbf{X}^H}(x) \end{aligned} \tag{6.30}$$

where these integrals are computed by using the convention $1/0 = \infty$ and $1/\infty = 0$. $out(\mathbf{X})^2$ is the second moment of singular value distribution of $\mathbf{X}$ and when $\mathbf{X}$ is invertible (or has no zero eigenvalues), $in(\mathbf{X})^2$ is the second moment of singular value distribution of $\mathbf{X}^{-1}$.

6.2.1 Classes of R-diagonal Matrices

A Haar-unitary matrix $\mathbf{V}$ and a (i.i.d.) Gaussian random matrix $\mathbf{X}$ are asymptotically R-diagonal matrices because they can be decomposed as

$$\mathbf{V} = \mathbf{VI}; \qquad \mathbf{X} = \mathbf{UQ} \tag{6.31}$$

where $\mathbf{U}$ is a Haar-unitary matrix and $\mathbf{Q}$ is a quarter circle distributed random matrix. Moreover, with the following theorems we here present some important class R-diagonal matrices:

Theorem 6.2.2 (Haagerup and Larsen (2000) [292]) Let the matrix $\mathbf{X}_i$ be a free family of R-diagonal matrices for all $1 \leq i \leq L$. Then

- sum of free R-diagonal matrices: $\mathbf{X}_1 + \cdots + \mathbf{X}_L = \sum_{i=1}^{L} \mathbf{X}_i$;
- product of free R-diagonal matrices: $\mathbf{X}_1 \cdots \mathbf{X}_L = \prod_{i=1}^{L} \mathbf{X}_i$;
- power of a R-diagonal matrices: $\left(\mathbf{X}_i\right)^p, i = 1, \ldots, L$ for a natural numbers p

are R-diagonal, too.

Theorem 6.2.3 (proposition 6.1.1 in [259]) Let the matrix $\mathbf{X}$ be R-diagonal and free of the matrix $\mathbf{Y}$. Then $\mathbf{XY}$ is also R-diagonal.

Theorem 6.2.4 ([259]) Let the free Hermitian matrices $\mathbf{X}$ and $\mathbf{Y}$ have a symmetric (even) eigen value distribution on the real line. Then the matrix $\mathbf{XY}$ is R-diagonal.

A large class of R-diagonal matrices have the property of behaving as if they are identical with respect to multiplication:

Theorem 6.2.5 (Proposition 3.10 in Haagerup and Larsen (2000) [292]) Let the random matrices $\mathbf{X}_i, i = 1, \ldots, L$ be asymptotically free R-diagonal elements and their asymptotic eigenvalue distributions of $\mathbf{X}_i$ be identical for all i. Then the asymptotic eigenvalue distribution of

$$\mathbf{X}_1 \cdots \mathbf{X}_L = \prod_{i=1}^{L} \mathbf{X}_i$$

is identical to $\mathbf{X}_i, i = 1, \ldots, L$.

6.2.2 Additive Free Convolution

Operations such as a sum or product of R-diagonal random matrices can be performed without quaternionic free calculus.

Consider a Hermitian matrix $\tilde{\mathbf{H}}$ such that the empirical eigenvalue distribution of $\tilde{\mathbf{H}}$ is

$$p_{\tilde{\mathbf{X}}}(x) = \frac{1}{2}\left[p_{\sqrt{\mathbf{X}\mathbf{X}^H}}(x) + p_{\sqrt{\mathbf{X}\mathbf{X}^H}}(-x)\right] \tag{6.32}$$

where $\tilde{\mathbf{H}}$ is symmetrized singular value version of $\mathbf{X}$.

Theorem 6.2.6 (Proposition 3.5 in [292]) Let the asymptotically free random matrices $\mathbf{A}$ and $\mathbf{B}$ be R-diagonal. Define

$$\mathbf{C} = \mathrm{A} + \mathrm{B}$$

Then we have

$$R_{\tilde{\mathbf{C}}}(\omega) = R_{\tilde{\mathbf{A}}}(\omega) + R_{\tilde{\mathbf{B}}}(\omega) \tag{6.33}$$

Two lemmas make the problem as trivial as in Hermitian case.

Lemma 6.2.7 (Symmetrization Lemma I) Let $\mathbf{X}$ be a rectangular non-Hermitian random matrix in general. Then we have,

$$G_{\tilde{\mathbf{X}}}(s) = sG_{\sqrt{\mathbf{X}\mathbf{X}^H}}\left(s^2\right) \tag{6.34}$$

Lemma 6.2.8 (Symmetrization Lemma II) Let the matrix $\mathbf{X}$ be defined as in the previous lemma. Then we have

$$S_{\tilde{\mathbf{X}}}(z) = \left[\frac{z+1}{z} S_{\mathbf{X}\mathbf{X}^H}(z)\right]^{1/2} \tag{6.35}$$

Example 6.2.9 (deformed quarter-circle element) Let $\mathbf{X}$ be a deformed quarter-circle element. Then find the R-transform $R_{\tilde{\mathbf{X}}}(\omega)$. Recall (5.154) such that the S-transform of $\mathbf{X}\mathbf{X}^H$ is given by

$$S_{\mathbf{X}\mathbf{X}^H}(z) = \frac{1}{z+c} \tag{6.36}$$

Using the inversion formula between R-transform and S-transform (5.148), we have

$$\omega\tilde{R}_{\mathbf{X}}(\omega)\left(\omega\tilde{R}_{\mathbf{X}}(\omega)+1\right)\cdot\frac{1}{\omega\tilde{R}_{\mathbf{X}}(\omega)+c}=\omega^2$$

which leads to the following solution

$$\tilde{R}_{\mathbf{X}}(\omega)=\frac{\omega}{2}-\frac{1}{2\omega}+\sqrt{\left(\frac{\omega}{2}-\frac{1}{2\omega}\right)^2+c} \tag{6.37}$$

Note that we are two solutions. In the case of $c=1$, the right one must give $\tilde{R}_{\mathbf{X}}(\omega)=\omega$. See Example 5.8.10 for more details about the approach of how to choose the right solutions from the two solutions. □

6.2.3 Multiplicative Free Convolution

The trace operator is cyclic invariant. This allows us to work on complex value-free probability by means of S-transform to deal with a multiplication of non-Hermitian matrices.

The following theorem gives a straightforward way to switch from singular values to eigenvalues of R-diagonal matrices and vice versa.

Theorem 6.2.10 ([292]) Let the random matrix $\mathbf{X}$ be R-diagonal such that it can be decomposed as $\mathbf{X}=UY$, where $\mathbf{U}$ is Haar unitary matrix and free of the matrix $\mathbf{Y}=\sqrt{\mathbf{X}\mathbf{X}^H}$. Then we have:

(i) The eigenvalue distribution $P_{\mathbf{X}}(z)$ is circularly invariant with its boundary

$$\operatorname{supp}\left(P_{\mathbf{X}}\right)=\left[\operatorname{in}(\mathbf{X})^{-1},\operatorname{out}(\mathbf{X})\right]\times_p[0,2\pi] \tag{6.38}$$

Here by $\times_p$ we denote the polar set product: $A\times_p B=\left\{ae^{j\theta}\,|a\in A,\theta\in B\right\}$. Explicitly, the support of the eigenvalue distribution$P_{\mathbf{X}}(z)$ is the annulus with the inner radius $\operatorname{in}(\mathbf{X})^{-1}$ and the outer radius $\operatorname{out}(\mathbf{X})$. (ii) The S-transform $S_{\mathbf{Y}^2}(z)$ of $\mathbf{Y}^2$ has an analytic continuation to the neighborhood of interval $\left(P_{\mathbf{Y}^2}(0)-1,0\right]$ and monotonically deceasing on $\left(P_{\mathbf{Y}^2}(0)-1,0\right]$ such that the derivative of the S-transform $S'_{\mathbf{Y}^2}(z)<0$, and it takes the values in between

$$S\left(\left(P_{\mathbf{Y}^2}(0)-1,0\right]\right)=\left(\operatorname{in}(\mathbf{X})^{-2},\operatorname{out}(\mathbf{X})^2\right] \tag{6.39}$$

(iii) $P_{\mathbf{X}}(z)\big|_{z=0}=P_{\mathbf{Y}}(z)\big|_{z=0}$ and the radial distribution function

$$P_{\mathbf{X}}^{(-1)}(r)\left(\frac{1}{\sqrt{S_{\mathbf{Y}^2}(r-1)}}\right)=r;\qquad r\in\left(P_{\mathbf{Y}}(0),1\right] \tag{6.40}$$

(iv) The eigenvalue distribution $P_{\mathbf{X}}(z)$ is the only circularly symmetric probability measure satisfying (iii).

Corollary 6.2.11 ([292]) With the notation as in Theorem 6.2.10, the functional inversion of radial probability measure of $\mathbf{X}$

$$P_{\mathbf{X}}^{(-1)}(r)=\frac{1}{\sqrt{S_{\mathbf{Y}^2}(r-1)}}:\left(P_{\mathbf{Y}}(0),1\right]\to\left[\operatorname{in}(\mathbf{X})^{-1},\operatorname{out}(\mathbf{X})\right] \tag{6.41}$$

has an analytical continuation to a neighborhood of its domain and monotonically increasing on $\left(P_{\mathbf{Y}}(0), 1\right]$ such that the derivative $dP_{\mathbf{X}}^{(-1)}(r)/dr > 0$. Moreover, the radial density of $\mathbf{X}$ such that

$$2\pi r p_{\mathbf{X}}(z)\Big|_{|z|=r} = \frac{dP_{\mathbf{X}}(r)}{dr}, \qquad r \in \left[\text{in}(\mathbf{X})^{-1}, \text{out}(\mathbf{X})\right] \tag{6.42}$$

has an analytical continuation to the neighborhood of $\left(\text{in}(\mathbf{X})^{-1}, \text{out}(\mathbf{X})\right]$.

Theorem 6.2.10 and its corollary play central to characterize non-Hermitian random matrices.

Example 6.2.12 (project compression) Let the entries of the $T \times T$ matrix $\mathbf{G}$ be independent identically distributed with variance $1/T$ and the matrix $\mathbf{P} \in \{0,1\}^{T\times T}$ be diagonal with K nonzero entries. Then, show that the empirical eigenvalue distribution of $\mathbf{H} = \mathbf{GP}$ converges almost surely to

$$p(z) = (1-\alpha)\delta(z) + \begin{cases} 1/\pi & ,|z| < \sqrt{a} \\ 0 & ,\text{elsewhere} \end{cases}$$

First, $\mathbf{HH}^H$ is the square equivalent of the deformed quarter circle law (eigenvalues) element. Thus

$$S_{\mathbf{HH}^H}(z) = \frac{1}{z+\alpha}$$

With Theorem 6.2.10, we have

$$P_{\mathbf{H}}^{(-1)}(r) = \frac{1}{\sqrt{S_{\mathbf{HH}^H}(r-1)}} = \sqrt{r+\alpha-1}$$

Then the probability measure (radial) is given by

$$P_{\mathbf{H}}(r) = (1-\alpha) + r^2$$

Moreover, the zero measure of the distribution is

$$P_{\mathbf{H}}(z)\big|_{z=0} = (1-\alpha)\delta(z)$$

where the asymptotic eigenvalue distribution of $\mathbf{H}$ can be easily obtained as

$$\begin{aligned} p_{\mathbf{H}}(z) &= (1-\alpha)\delta(z) + \left(\tfrac{1}{2\pi r}\tfrac{dP_{\mathbf{H}}(r)}{dr}\right)\Big|_{|z|=r} \\ &= (1-\alpha)\delta(z) + \tfrac{1}{\pi} \end{aligned}$$

Finally, we need determine the boundary of the density. Since the distribution has some zero measure, then inner radius of the density is given by

$$\text{in}(\mathbf{X})^{-1} = 0$$

The outer radius of the density reads

$$\text{out}(\mathbf{X}) = \frac{1}{\sqrt{S_{\mathbf{HH}^H}(z)}}\Big|_{z=0} = \sqrt{\alpha}$$

□

Theorem 6.2.13 (Proposition 3.10 in Haagerup and Larsen (2006) [292]) Let the random matrices $\mathbf{X}_k$ be asymptotically free R-diagonal elements, and their asymptotic eigenvalue distributions of $\mathbf{X}_k$ be identical for all $k = 1, 2, \ldots, N$. Then the asymptotic eigenvalue distributions of

$$\prod_{k=1}^{N} \mathbf{X}_k \tag{6.43}$$

and $\mathbf{X}_k^N$ are identical for any $k = 1, 2, \ldots, N$.

Theorem 6.2.13 has a lot of practical applications.

The R-diagonal matrices has an interesting consequence of additively free convolution regarding to singular values:

Theorem 6.2.14 ([293]) Let $\mathbf{X}$ be R-diagonal matrix, and has the decomposition form $\mathbf{X} = \mathbf{U}\sqrt{\mathbf{Y}_1}$ such that $\mathbf{U}$ is Haar-unitary matrix and free of the matrix $\sqrt{\mathbf{Y}_1} = \sqrt{\mathbf{X}\mathbf{X}^H}$. Furthermore, let the asymptotic eigenvalue distribution of $\mathbf{X}$ be

$$\alpha\delta(z) + \begin{cases} p_{\mathbf{X}}(z) & P_{\mathbf{X}}^{(-1)}(\alpha) < |z| \leqslant b \\ 0 & \text{elsewhere} \end{cases} \tag{6.44}$$

Moreover, define a summation of identical free matrices as

$$\mathbf{Y}_c = \sum_{k=1}^{1/c} \mathbf{Y}_k$$

for $c \leq 1$. Then, the asymptotic eigenvalue distribution of $\mathbf{X}_c = \mathbf{U}\sqrt{\mathbf{Y}_c}$ satisfies

$$p_{\mathbf{X}_c}(z) = \alpha_c\delta(z) + \begin{cases} p_{\mathbf{X}}\left(\frac{z}{\sqrt{c}}\right) & \sqrt{c}P_{\mathbf{X}}^{(-1)}\left(\frac{1}{c}\alpha_c + 1 - \frac{1}{c}\right) < |z| \leqslant \sqrt{c}P_{\mathbf{X}}^{(-1)}(1) \\ 0 & \text{elsewhere} \end{cases}$$

where $\alpha_N = \max(0, 1 + cN - N)$.

Theorem 6.2.14 immediately inspires us to propose the following theorem:

Theorem 6.2.15 (Theorem 29 of Cakmak and Muller (2012) [139]) Consider $N \times N$ R-diagonal matrix $\mathbf{X} = \left[\mathbf{x}_1, \ldots, \mathbf{x}_N\right]$ whose eigenvalue distribution is

$$\alpha\delta(z) + \begin{cases} p_{\mathbf{X}}(z) & P_{\mathbf{X}}^{(-1)}(\alpha) < |z| \leqslant b \\ 0 & \text{elsewhere} \end{cases}$$

where $P_{\mathbf{X}}^{(-1)}(r)$ is the functional inversion of radial probability measure (CDF). Moreover, let the $N \times T$ ($T \leq N$) matrix $\mathbf{X}_c = \left[\mathbf{x}_1, \ldots, \mathbf{x}_T\right]$, for $c = T/N \leq 1$. Define

$$\alpha_c = \max\left(0, 1 + \frac{\alpha}{c} - \frac{1}{c}\right)$$

Then, the empirical eigenvalue distribution of $\mathbf{X}_{c,u}$ such that

$$\mathbf{X}_{c,u} = \mathbf{U}\sqrt{\mathbf{X}_c^H\mathbf{X}_c}$$

converges almost surely to limit distribution satisfies

$$p_{\mathbf{X}_{c,u}}(z) = \alpha_c \delta(z) + \begin{cases} \frac{1}{c} p_{\mathbf{X}}(z) \; P_{\mathbf{X}}^{(-1)}\left(c\alpha_c + 1 - c\right) < |z| \leqslant b \\ 0 \qquad\qquad\qquad \text{elsewhere} \end{cases}$$

as $N, T \to \infty$ with $c = T/N \leq 1$ fixed.

Proof. We follow [139] for this proof.

$$\mathbf{X} = \left[\mathbf{x}_1, \ldots, \mathbf{x}_N\right]$$

Define an $N \times N$ diagonal matrix $\mathbf{P}$ whose diagonal terms are distributed as

$$p_{\mathbf{P}}(x) = (1-c)\,\delta(x) + c\delta(x-1).$$

Then, using the matrix projection, we have

$$\mathbf{P}\mathbf{X}\mathbf{X}^H\mathbf{P} = \mathbf{X}_c\mathbf{X}_c^H \tag{6.45}$$

which gives

$$p_{\mathbf{X}_c\mathbf{X}_c^H}(x) = (1-c)\,\delta(x) + c p_{\mathbf{X}_c\mathbf{X}_c^H}(x-1) \tag{6.46}$$

With Theorem 5.8.13(Theorem 14.10 in [259]), we have

$$R_{\mathbf{X}_c^H\mathbf{X}_c}(\omega) = R_{\mathbf{X}\mathbf{X}^H}(c\omega) \tag{6.47}$$

Recall the functional relation between R-transform and S-transform [259]

$$zR(z)\,S(zR(z)) = z; \qquad zS(z)\,R(zS(z)) = z \tag{6.48}$$

Using (6.48), we obtain

$$\begin{aligned} S_{\mathbf{X}_c^H\mathbf{X}_c}(z) &= \frac{1}{R_{\mathbf{X}_c^H\mathbf{X}_c}\left(zS_{\mathbf{X}_c^H\mathbf{X}_c}(z)\right)} = \frac{1}{R_{\mathbf{X}\mathbf{X}^H}\left(czS_{\mathbf{X}_c^H\mathbf{X}_c}(z)\right)} \\ &= S_{\mathbf{X}\mathbf{X}^H}\left(czS_{\mathbf{X}_c^H\mathbf{X}_c}(z)\,R_{\mathbf{X}\mathbf{X}^H}\left(zS_{\mathbf{X}_c^H\mathbf{X}_c}(z)\right)\right) \\ &= S_{\mathbf{X}\mathbf{X}^H}(cz) \end{aligned} \tag{6.49}$$

Moreover we defined in the Theorem

$$\mathbf{X}_{c,u} = \mathbf{U}\sqrt{\mathbf{X}_c^H\mathbf{X}_c} \tag{6.50}$$

With (6.41) , we have

$$\begin{aligned} P_{\mathbf{X}}^{(-1)}(r) &= \frac{1}{\sqrt{S_{\mathbf{X}_c^H\mathbf{X}_c}(r-1)}} = \frac{1}{\sqrt{S_{\mathbf{X}\mathbf{X}^H}(cr-c)}} \\ &= \frac{1}{\sqrt{S_{\mathbf{X}\mathbf{X}^H}((cr+1-c)-1)}} \\ &= P_{\mathbf{H}}^{(-1)}(cr+1-c) \end{aligned} \tag{6.51}$$

Let $r \to P_{\mathbf{X}_{c,u}}(r)$. We have

$$P_{\mathbf{X}}(r) = cP_{\mathbf{X}_{c,u}}(r) + 1 - c \tag{6.52}$$

so

$$P_{\mathbf{X}_{c,u}}(r) = \frac{1}{c}P_{\mathbf{X}}(r) + 1 - \frac{1}{c}$$

Moreover, zero measure can be easily found as

$$\begin{aligned}\alpha_c &= P_{\mathbf{X}_{c,u}}(r)\Big|_{r=0} = \max\left(0, 1-\tfrac{1}{c}+\tfrac{1}{c}P_{\mathbf{X}}(r)\big|_{r=0}\right)\\ &= \max\left(0, \tfrac{\alpha}{c}+1-\tfrac{1}{\alpha}\right),\end{aligned} \tag{6.53}$$

since we defined zero measure of $\mathbf{X}$ as α. Thus, we have distribution of $\mathbf{X}_{c,u}$, which satisfies with

$$\begin{aligned}\tfrac{dP_{\mathbf{X}_{c,u}}(r)}{dr} &= p\mathbf{X}_{c,u}(r)\\ &= \alpha_c\delta(r)+\tfrac{1}{c}p_{\mathbf{X}}(r)\\ p_{\mathbf{X}_{c,u}}(z) &= \tfrac{1}{2\pi r}p\mathbf{X}_{c,u}(r)\Big|_{r=|z|}\\ &= \alpha_c\delta(z)+\tfrac{1}{c}p_{\mathbf{X}}(z)\end{aligned}$$

In the final step, we study how the boundary of the distribution function is affected by the matrix transform (6.50). It is obvious that the outer boundary does not change since (with (6.51))

$$P^{(-1)}_{\mathbf{X}_{c,u}}(r)\big|_{r=1} = P^{(-1)}_{\mathbf{X}_{c,u}}(cr+r-c)\big|_{r=1} = P^{(-1)}_{\mathbf{X}}(r)\big|_{r=1}.$$

In the same way, the inner boundary is given by

$$P^{(-1)}_{\mathbf{X}_{c,u}}(r)\Big|_{r=\alpha_c} = P^{(-1)}_{\mathbf{X}_{c,u}}(cr+1-c)\Big|_{r=\alpha_c}$$

Thus, the asymptotic eigenvalue distribution of $\mathbf{X}_{c,u}$ converges to the limit distribution

$$p_{\mathbf{X}_{c,u}}(z) = \alpha_c\delta(z)+\begin{cases}\tfrac{1}{c}p_{\mathbf{X}}(z)\ P^{(-1)}_{\mathbf{X}_{c,u}}(cr+1-c)\Big|_{r=\alpha_c} < |z| \leqslant b\\ 0 \qquad\qquad\qquad\qquad \text{elsewhere,}\end{cases}$$

as $N, T \to \infty$ with the ratio $c = T/N \leq 1$ fixed. □

The above proof can be viewed as an example to illustrate the approach. For N time series of length T, we can model these time series using an $N \times T$ non-Hermitian random matrix $\mathbf{X} \in \mathbb{C}^{N\times T}$ where N and T are large, in the order of 100–5000. The current laptop can handle a matrix of $5000 - 5000$ for their eigenvalues calculations.

6.2.4 Isotropic Random Matrices

In the physics literature, the R-diagonal matrix is also called *isotropic random matrix*. Here we give an intuitive introduction to this concept. The concept of isotropic random matrices in analogy to isotropic complex random variables z that have a circularly symmetric probability distribution depending only on the module $|z|$. Using polar decomposition, one can write $z \equiv re^{j\phi}$ where r is a real non-negative random variable and ϕ is a random variable (phase) with a uniform distribution on $[0, 2\pi)$.

Isotropic random matrices are defined by a straightforward generalization of isotropic complex random variables. A square $N \times N$ matrix $\mathbf{X}$ is said to be an isotropic random matrix if it has a polar decomposition $\mathbf{X} = \mathbf{HU}$ in which $\mathbf{H}$ is a positive semidefinite Hermitian random matrix and $\mathbf{U}$ is a unitary random matrix independent of $\mathbf{H}$ and distributed on the unitary group $\mathcal{U}(N)$ with the Haar measure. In short, $\mathbf{U}$ is a Haar unitary matrix. In other words, for an $N \times N$ isotropic random matrix $\mathbf{X}$, one has $P(\mathbf{X}) = P(\mathbf{XV})$, where $\mathbf{V} \in \mathcal{U}(N)$.

Examples of isotropic random matrices include: (i) Girko-Ginibre matrix; (ii) the matrix of the form **UHV** where **U** and **V** are unitary Haar measure random matrices and **H** is a positive semidefinite Hermitian random matrix.

Properties of for isotropic random matrices include: (i) Eigenvalue spectrum is rotationally symmetric on the complex plane; (ii) They form isotropic unitary ensemble (IUE). (iii) Average eigenvalue distribution for the product of L matrices generated from any type of IUE is independent of the order of multiplication in the $N \to \infty$ limit.

6.3 The Sum of Non-Hermitian Random Matrices

We define the arithmetic average of L matrices as

$$\frac{1}{L}\left(\mathbf{X}_1 + \cdots + \mathbf{X}_L\right) = \frac{1}{L}\sum_{i=1}^{L}\mathbf{X}_i \tag{6.54}$$

where $\mathbf{X}_i \in \mathbb{C}^{N\times n}, i = 1, \ldots, L$ are non-Hermitian random matrices, whose entries are i.i.d. with zero mean and variance one.

It is found that the sum will not affect the empirical eigenvalue density function, which converges to the Marchenko–Pastur law (Theorem 3.6.1) for the large matrix limit. In other words, the matrix sum affects the scaling parameter σ^2 only if the Marchenko–Pastur law is found to be identical to L. Recall from Theorem 3.6.1 that the eigenvalues of $\frac{1}{n}\mathbf{X}\mathbf{X}^H$ or the singular values of $\mathbf{X}$ have the probability distribution function (PDF) defined as

$$f_{MP}(x) = \frac{1}{2\pi x c\sigma^2}\sqrt{(b-x)(x-a)}\,\mathbb{I}(a \leqslant x \leqslant b) \tag{6.55}$$

where

$$a = \sigma^2\left(1-\sqrt{c}\right)^2, \quad b = \sigma^2\left(1+\sqrt{c}\right)^2, \text{ and } c = n/N \tag{6.56}$$

Consider the $N \times T$, $(T \leq N)$ matrix $\mathbf{X}$ and the $N \times (N-T)$ null matrix $\mathcal{N}$. Let the $N \times N$ matrix $\mathbf{X}_s = [\mathbf{X}\,|\mathbf{N}]$. Then we have

$$\mathbf{X}_s\mathbf{X}_s^H = \mathbf{X}\mathbf{X}^H \tag{6.57}$$

We call $\mathbf{X}_s$ the square equivalence of $\mathbf{X}$. When the matrix is R-diagonal that also means that the matrix is biunitarily invariant, then square equivalence can be replaced by square equivalent. Let the $N \times N$ random matrix $\mathbf{X}$ be R-diagonal such that

$$\mathbf{X}_\beta = \left[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\right]$$

with the ratio $\beta = T/N \leq 1$ fixed. Moreover, define a $N \times T$ random matrix as

$$\mathbf{X} = \left[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\right]$$

Define an arbitrary $N \times N$ diagonal matrix $\mathbf{P}$ such that the diagonal entries

$$p_{\mathbf{P}}(x) = (1-\beta)\,\delta(x) + \beta\delta(x-1)$$

Using $\mathbf{X}\mathbf{P} = \mathbf{X}_p$, then we have

$$\mathbf{X}_p\mathbf{X}_p^H \equiv \mathbf{X}_\beta\mathbf{X}_\beta^H \tag{6.58}$$

Thus we call $\mathbf{X}_p$ the square equivalent of a rectangular random matrix $\mathbf{X}_\beta$.

Example 6.3.1 (matrix-valued hypothesis testing) Consider the hypothesis testing

$$\begin{aligned} \mathcal{H}_0 &: \mathbf{X}_1 + \cdots + \mathbf{X}_L \\ \mathcal{H}_1 &: \left(\mathbf{X}_1 + \cdots + \mathbf{X}_L\right) + \left(\mathbf{Y}_1 + \cdots + \mathbf{Y}_K\right) \end{aligned} \tag{6.59}$$

where $\mathbf{X}_i \in \mathbb{C}^{N\times n}, i = 1, \ldots, L$ are non-Hermitian random matrices, whose entries are i.i.d. with zero mean and variance one. Here $\mathbf{Y}_i \in \mathbb{C}^{N\times n}, i = 1, \ldots, K$ are non-Hermitian random matrices. $\mathbf{Y}_i \in \mathbb{C}^{N\times n}, i = 1, \ldots, K$ are freely independent of $\mathbf{X}_i \in \mathbb{C}^{N\times n}, i = 1, \ldots, L$.

Then the singular values of $\mathbf{X}_{\text{sum}} = \mathbf{X}_1 + \cdots + \mathbf{X}_L$ the probability distribution function defined in (3.10) with the variance parameter $\sigma^2_{\text{sum}} = \sigma^2_1 + \cdots + \sigma^2_L$. Using $\tilde{\mathbf{Z}}$ to represent the square equivalence of $\mathbf{Z}$ defined in (6.57) or the square equivalent of $\mathbf{Z}$ (6.58), with the aid of the linearity of the trace function, we have

$$\mathcal{H}_0 : \text{Tr}\left(\tilde{\mathbf{X}}_{\text{sum}}\right)$$

$$\mathcal{H}_1 : \text{Tr}\left(\tilde{\mathbf{X}}_{\text{sum}}\right) + \text{Tr}\left(\tilde{\mathbf{Y}}_1 + \cdots + \tilde{\mathbf{Y}}_K\right)$$

which is the standard scalar-valued hypothesis-testing problem. □

Figures 6.2 and 6.3 illustrate the above remarks. The average operation will not affect the PDF of $\mathbf{X}$, if we have L realizations of $\mathbf{X}$, i.e., $\mathbf{X}_i, i = 1, \ldots, L$.

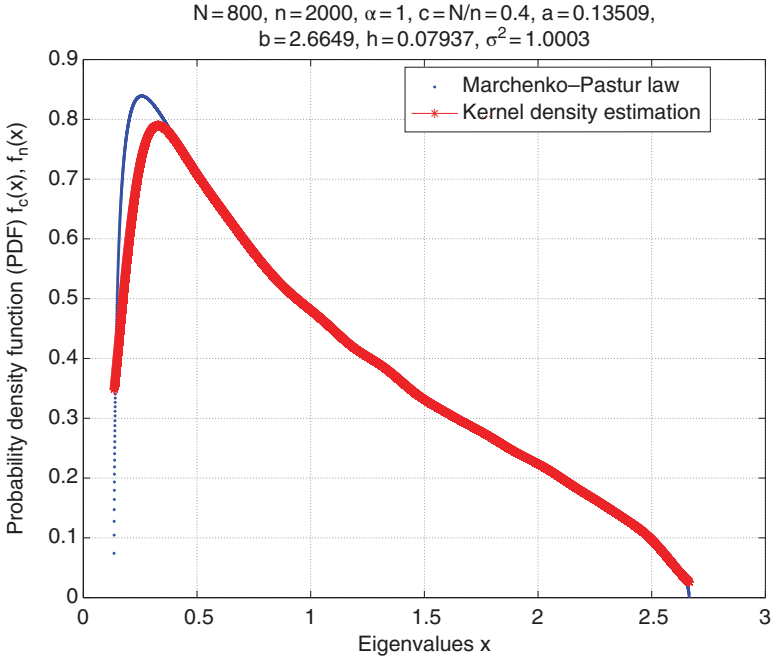


Figure 6.2 The sum of L non-Hermitian random matrices: $N = 800, n = 2000, c = N/n = 0.4$, $a = 0.135, b = 2.66$ and $h = 0.079$, for one matrix, i.e., $L = 1$.

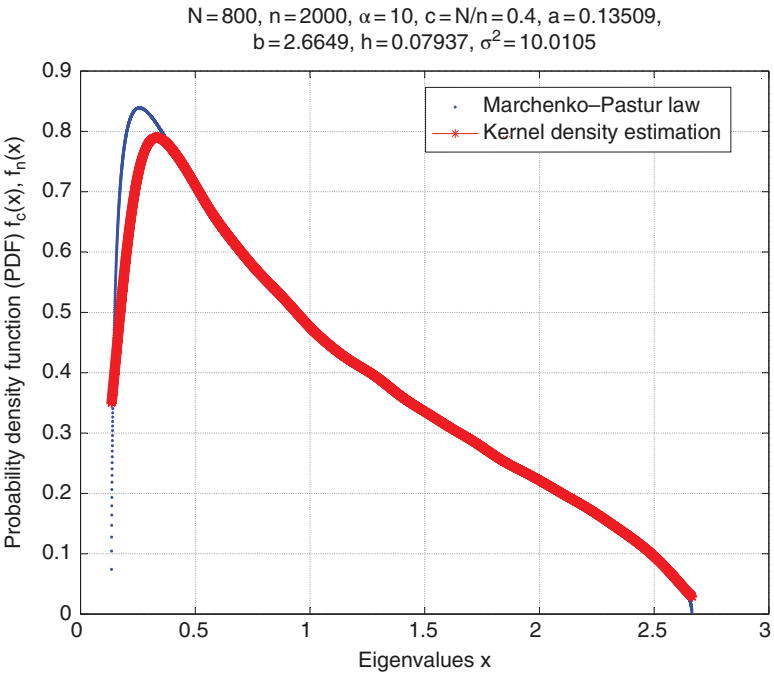


Figure 6.3 The same as Figure 6.2 except $L = 10$.

MATLAB Code: Sum of Non-Hermitian Matrices

```
clear all;
%Reference
% Non-Hermitian Random Matrix Theory for MIMO Channels
% Burak Cakmak (2012) MS Thesis
% NTNU-Trondheim Norwegain University of Science
  and Technology
N=200*4; beta=0.4; kappa=0.05; alpha=1; c=beta; n=N/c;
% c=N/n; c=p/n; beta=T/R=1/c; beta>1.
R=N; T=n; rho=beta/kappa; S=R*rho; %kappa=beta/rho
radius_inner=((1-kappa)*(1-beta))^(alpha/2) sigma=1;
 step=0.01/40; %step=0.01/10/4/2/2;
h=1/n^(1/3);
a=(1-sqrt(c))^2; b=(1+sqrt(c))^2;
x=(a+step):step:b;
fcx=(1/2/pi/c./x).*sqrt((b-x).*(x-a));
% the density function of Marcenko and Pastur law


Z=zeros(N,n);
for i=1:alpha
```

```
Y=randn(N,n)
+sqrt(-1)*randn(N,n);
 % N x n matrix white noise that follows the
   Marchenko-Pastur Law


Z=Z+Y; % singular value equivalent
end % i

VarZ=var(Z)';
VarZ(1:1)

for j=1:n
Z(:,j)=Z(:,j)/std(Z(:,j)); % normalized the variance to one
end %j

lambda=eig(1/n*Z*Z'); % eigenvalues of sample covariance
matrix

L=(b-a)/step;
x1=a+step;
for j=1:L

for i=1:N y=(x1-lambda(i))/h;
Ky(i)=kernel(y);
end %N
fnx(j)=sum(Ky)/N/h;
x1=x1+step;
x2(j)=x1;
end %L

% figures
ifig=0;

ifig=ifig+1;figure(ifig)
plot(x,fcx,'.b', x2,fnx,'-*r')
xlabel('Eigenvalues x')
ylabel('Probability Density
Function (PDF) f_c(x), f_n(x)')
 legend('Marcenko-Pastur Law','Kernel Density Estimation');
title(['N=',int2str(N),', n=',int2str(n),',
\alpha=',num2str(alpha), .. ', c=N/n=',num2str(c), ',
a=',num2str(a), ', b=',num2str(b),', h=',num2str(h)]) grid

function [Kx] = kernel(x)
Kx=1/sqrt(2*pi)*exp(-0.5*x.^2);
```

6.4 The Product of Non-Hermitian Random Matrices

We define the product as

$$\mathbf{X}_1 \cdots \mathbf{X}_L = \prod_{i=1}^{L} \mathbf{X}_i$$

where $\mathbf{X}_i \in \mathbb{C}^{N\times n}, i = 1, \ldots, L$ are non-Hermitian random matrices, whose entries are i.i.d. with zero mean and variance one. Similar to (6.54) for the arithmetic average, we are motivated to understand the geometric mean of L non-Hermitian random matrices:

$$\left(\mathbf{X}_1 \cdots \mathbf{X}_L\right)^{1/L} = \left(\prod_{i=1}^{L} \mathbf{X}_i\right)^{1/L}$$

The product of large non-Hermitian random matrices is much more involved than the sum. But we still have the tractable calculus for this kind of operations.

In Figure 6.4 and Figure 6.5, the case of $L = 1$ and $L = 10$ are compared for the eigenvalues on the complex plane. The so-called the single-ring law is observed: All eigenvalues lie within a single ring. The radius of the inner circle is greatly reduced when the number of matrices L increases.

Now we address how to calculate the radius of the inner circle and the probability distribution of the eigenvalues within the ring.

Definition 6.4.1 Consider the $N \times n$ matrix $\mathbf{X}$. Let the $n \times n$ matrix $\mathbf{U}$ be Haar-unitary matrix and free of $\mathbf{X}^H\mathbf{X}$ Moreover define,

$$\mathbf{X}_u = \mathbf{U}\sqrt{\mathbf{X}^H\mathbf{X}} \tag{6.60}$$

Then, as far as concerning singular value distribution, we have

$$\mathbf{X}_u^H\mathbf{X}_u \equiv \mathbf{X}^H\mathbf{X}, \quad \mathbf{X}^H\mathbf{X} \in \mathbb{C}^{n\times n} \tag{6.61}$$

The matrix $\mathbf{X}_u$ is called the **singular value equivalent** of $\mathbf{X}$.

Theorem 6.4.2 Let the entries of the $N \times n$ matrix $\mathbf{X}$ be i.i.d. with zero mean variance $1/N$. Then, the empirical eigenvalue distribution of the **singular value equivalent** of $\mathbf{X}$ converges almost surely to

$$f_{\mathbf{X}_u}(z) = \begin{cases} \frac{1}{c\pi} & \sqrt{1-c} < |z| \leqslant 1 \\ 0 & \text{elsewhere} \end{cases} \tag{6.62}$$

as $N, n \to \infty$ with the ratio $c = n/N \leq 1$ fixed.

The radius of the inner circle is $\sqrt{1-c}$, which agrees with the empirical eigenvalue distribution, as shown in Figure 6.4.

Now we consider the case

$$\prod_{i=1}^{L} \mathbf{X}_{u,i} \tag{6.63}$$

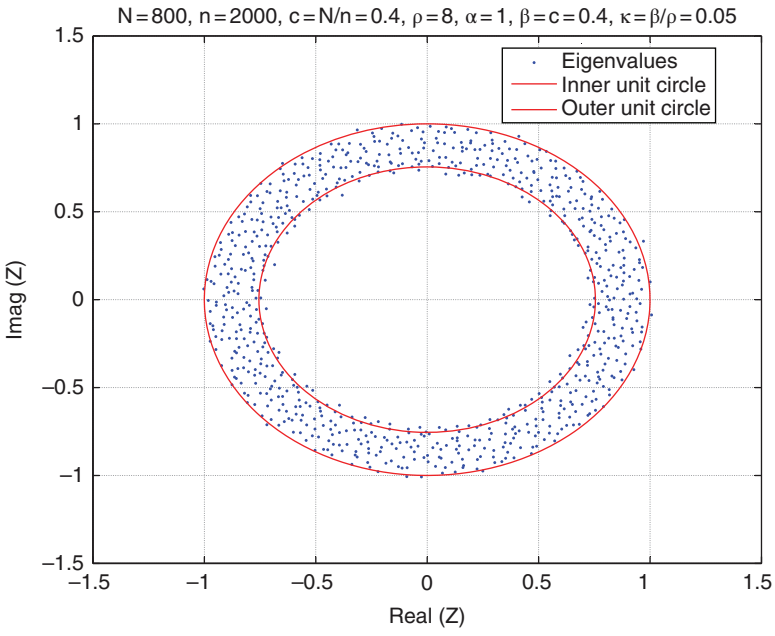


Figure 6.4 Eigenvalues for a product of L non-Hermitian random matrices: $N = 800, n = 2000, c = N/n = 0.4, a = 0.135, b = 2.66$ and $h = 0.079$, for one matrix, i.e., $L = 1$.

where $\mathbf{X}_{u,i}$ is the singular value equivalent of the rectangular $N \times n$ matrix $\mathbf{X}_i$, whose entries are i.i.d. with zero mean and variance $1/N$.

Theorem 6.4.3 Let $N \times n$ matrix $\prod_{i=1}^{L} \mathbf{X}_{u,i}$ be defined in (6.63). Then, the empirical eigenvalue distributions of $\prod_{i=1}^{L} \mathbf{X}_{u,i}$ converge almost surely to the same limit given by

$$f_{\prod_{i=1}^{L} \mathbf{X}_{u,i}}(z) = \begin{cases} \frac{1}{\pi c L}|z|^{2/L-2} & (1-c)^{L/2} \leqslant |z| \leqslant 1 \\ 0 & \text{elsewhere} \end{cases}$$

as $N, n \to \infty$ with the ratio $c = N/n \leq 1$ fixed.

The radius of the inner circle is $(1-c)^{L/2}$, which agrees with the empirical eigenvalue distribution, as shown in Figure 6.5 for $L = 10$. The PDF is also given above. Figure 6.6 and Figure 6.7 illustrates the case for $L = 1$ and $L = 10$.

Moreover, there is an interesting measure for square non-Hermitian random matrices; such a measure, called a left-right eigenvector correlation in the literature, tells how many pairs of eigenvalues are lying close in the complex plane. Consider the $N \times N$ non-Hermitian random matrix $\mathbf{X}$ with eigenvalue decomposition

$$\mathbf{X} = \mathbf{V}\mathbf{\Lambda}\mathbf{W}^{-1} = \sum_i \lambda_i \mathbf{v}_i \mathbf{w}_i^H$$

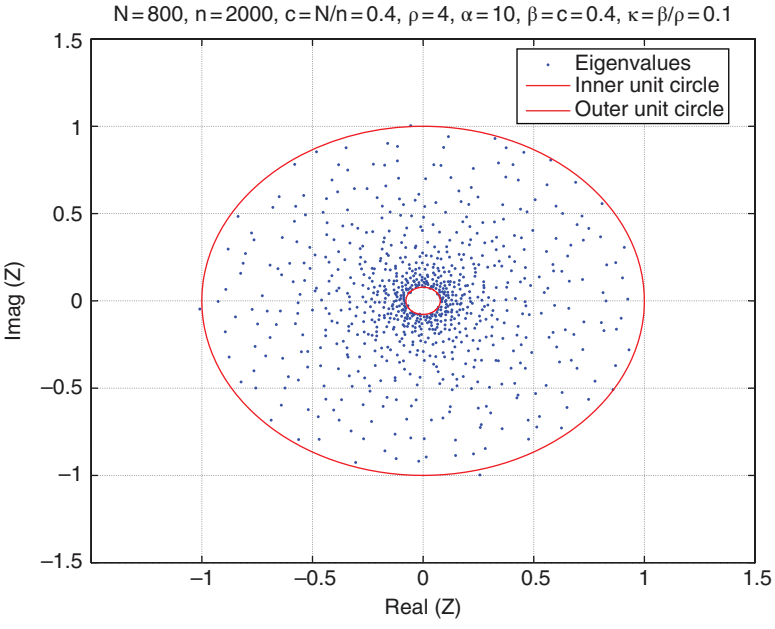


Figure 6.5 The same as Figure 6.4 except $L = 10$.

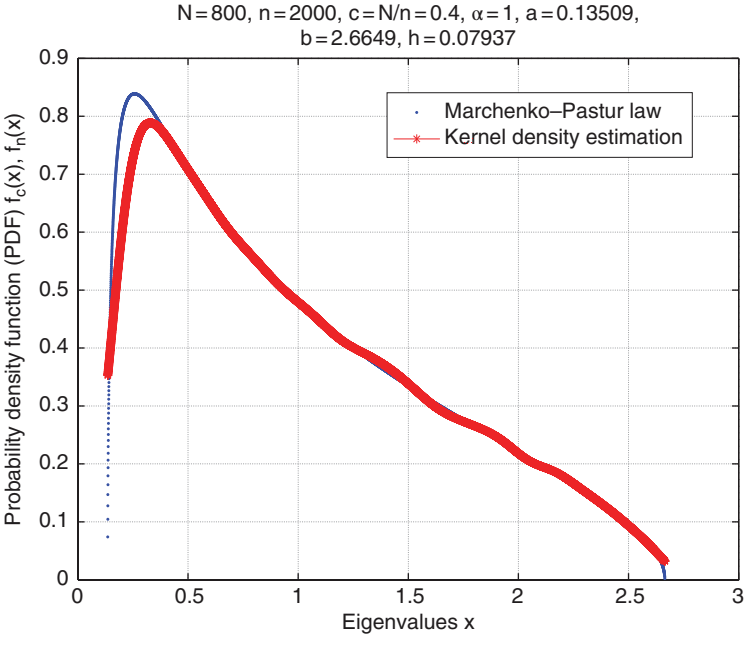


Figure 6.6 The empirical eigenvalue density function for a product of L non-Hermitian random matrices: $N = 800$, $n = 2000$, $c = N/n = 0.4$, $a = 0.135$, $b = 2.66$ and $h = 0.079$, for one matrix, i.e., $L = 1$.

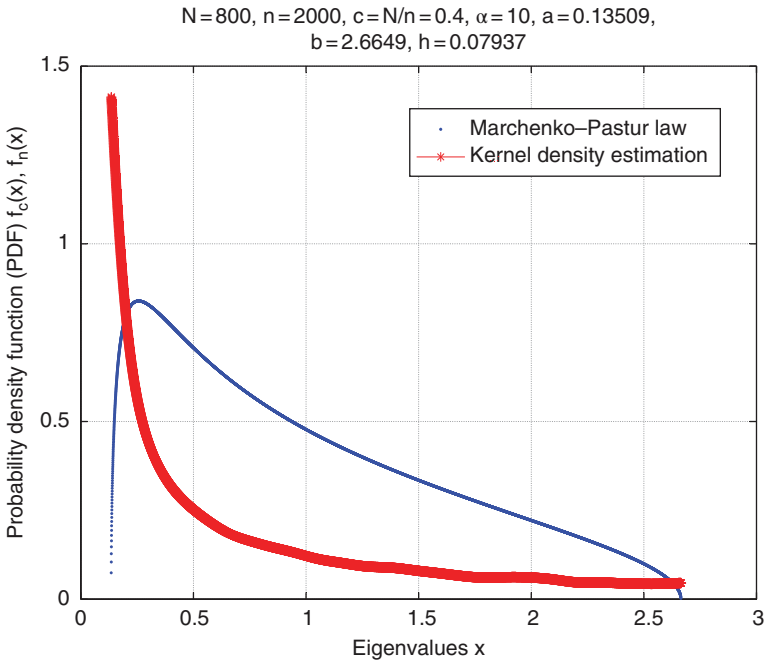


Figure 6.7 The same as Fig. 6.6 except $L = 10$.

where $\mathbf{V}$ is called right eigenvector matrix and $\mathbf{W} = \mathbf{V}^{-1}$ is called left eigenvectors. Then the correlation between right-left eigenvectors is defined as

$$C_{\mathbf{X}}(z) = \frac{\pi}{N} \sum_{i=1}^{N} \left(\mathbf{w}_i^H \mathbf{w}\right) \left(\mathbf{v}_i^H \mathbf{v}\right) \delta\left(z - z_i\right) \tag{6.64}$$

Theorem 6.4.4 Let the random matrix $\mathbf{X}$ defined in (6.63). Then the correlation between right-eigenvector defined in (6.64) of singular value equivalent of $\mathbf{X}$ is

$$C_{\mathbf{X}_u}(z) = \begin{cases} \frac{1}{c}\left(1 - |z|^{2/L}\right)|z|^{\frac{2}{L}-2} & (1-c)^{L/2} < |z| \leqslant 1 \\ 0 & \text{elsewhere} \end{cases} \tag{6.65}$$

as $N, n \to \infty$ with the ratio $c = N/n \leq 1$.

MATLAB Code: Product of Non-Hermitian Matrices

```
clear all;
%Reference
% Non-Hermitian Random Matrix Theory for MIMO Channels
% Burak Cakmak (2012) MS Thesis
% NTNU-Trondheim Norwegain University of Science
and Technology
N=200*4; beta=0.4; kappa=0.1; alpha=10; c=beta; n=N/c;
% c=N/n; c=p/n; beta=T/R=c.
```

```
R=N; T=n; rho=beta/kappa; S=R*rho; %kappa=beta/rho
radius_inner=(1-beta)^(alpha/2)
%radius_inner=((1-kappa)*(1-beta))^(alpha/2)
%radius_inner=((rho-beta)*(1-beta)*rho)^(alpha/2)
sigma=1;
 step=0.01/40; %step=0.01/10/4/2/2;
h=1/n^(1/3);
a=(1-sqrt(c))^2; b=(1+sqrt(c))^2;
x=(a+step):step:b;
fcx=(1/2/pi/c./x).*sqrt((b-x).*(x-a)); % the density
function of Marcenko and Pastur law
H=bernoulli(0.5,N,N)+sqrt(-1)*bernoulli(0.5,N,N);
% i.i.d. complex matrix
U=H*sqrtm(inv(H'*H)); % Unitrary Haar matrix U of N x N

Z=eye(N,N);
for i=1:alpha
H=1/sqrt(2)*randn(R,T)+sqrt(-1)*1/sqrt(2)*randn(R,T);
Z=Z*U*sqrtm(H*H'); % singular value equivalent
end % i

VarZ=var(Z)';
VarZ(1:10)

for j=1:N
Z(:,j)=Z(:,j)/std(Z(:,j)); % normalized the variance to one
end %j

Z=Z/sqrt(N); % normalized so the eigenvalues lie within
               unit circle

lambda=eig(Z*Z'); % eigenvalues of sample covariance matrix

% kernel density estimation
L=(b-a)/step;
x1=a+step;
for j=1:L
 for i=1:N
   y=(x1-lambda(i))/h;
   Ky(i)=kernel(y);
 end %N
 fnx(j)=sum(Ky)/N/h;
 x1=x1+step;
 x2(j)=x1;
end %L

% figures
```

```
ifig=0;

ifig=ifig+1;figure(ifig)
hist(lambda);
xlabel('Eigenvalues x')
ylabel('Probability Density Function (PDF) f(x)')
 legend('Kernel Density Estimation');
title(['N=',int2str(N),', n=',int2str(n),',
c=N/n=',num2str(c),',
a=',num2str(a),\ldots
 ', b=',num2str(b),', h=',num2str(h)])
grid

ifig=ifig+1;figure(ifig)
plot(x,n*h*(fcx-fnx))
xlabel('Eigenvalue x')
ylabel('Probability Density Function (PDF) ')
legend('Deviation n*h*[ f_c(x)-f_n(x) ]');
title(['N=',int2str(N),', n=',int2str(n),',
c=N/n=',num2str(c),',
a=',num2str(a),\ldots
 ', b=',num2str(b),', h=',num2str(h)])
grid

ifig=ifig+1;figure(ifig)
plot(x,fcx,'.b', x2,fnx,'-*r')
xlabel('Eigenvalues x')
ylabel('Probability Density
Function (PDF)
f_c(x), f_n(x)')
 legend('Marcenko-Pastur Law','Kernel Density Estimation');
title(['N=',int2str(N),', n=',int2str(n),',
c=N/n=',num2str(c),\ldots
', \alpha=',num2str(alpha), ', a=',num2str(a), ',
b=',num2str(b),\ldots ', h=',num2str(h)]) grid

ifig=ifig+1;figure(ifig); lambdaZ=eig(Z);
t=0:2*pi/1000:2*pi;x=sin(t);y=cos(t); % unit circle
plot(real(lambdaZ),imag(lambdaZ),'.',radius_inner*x,
radius_inner*y,'r-',x,y,'r-');
 axis([-1.5 1.5 -1.5 1.5])
%xlabel('Eigenvalues x')
xlabel('real(Z)'); ylabel('imag(Z)');
legend('Eigenvalues','Inner
Unit Circle ','Outer Unit Circle');
title(['N=',int2str(N),', n=',int2str(n),\ldots
 ', c=N/n=',num2str(c),', \rho=',num2str(rho),',
```

```
 \alpha=',num2str(alpha),\ldots
 ', \beta=c=',num2str(beta),
 ', \kappa=\beta/\rho=',num2str(kappa)])
grid

function [Kx] = kernel(x)
Kx=1/sqrt(2*pi)*exp(-0.5*x.^2);

function B=bernoulli(p,m,n);
% BERNOULLI.M
% This function generates n independent draws of a Bernoulli
% random variable with probability of success p.
% first, draw n uniform random variables

M = m;
N = n;
p = p;
B = rand(M,N) < p;
B=B*(-2)+ones(M,N);
```

6.5 Singular Value Equivalent Models

We consider the data matrix $\mathbf{Z}$ that can be expressed in the form

$$\mathbf{Z} = \mathbf{Y}_2\mathbf{Y}_1, \qquad \mathbf{Z} \in \mathbb{C}^{N\times n} \tag{6.66}$$

where $\mathbf{Y}_1 \in \mathbb{C}^{K\times n}, \mathbf{Y}_2 \in \mathbb{C}^{N\times K}$ are non-Hermitian random matrices. The entries of $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are assumed to be i.i.d. with zero mean and variance $1/N$, and $1/n$, respectively. The data matrix model defined in (6.66) generalizes the standard model in the previous sections. As the $\rho = K/N$ goes to infinity, i.e., $\rho \to \infty$

$$\lim_{\rho\to\infty} \mathbf{Z} \equiv \mathbf{X} \tag{6.67}$$

and the entries of $\mathbf{X} \in \mathbb{C}^{N\times n}$are i.i.d. with zero mean and variance $1/N$. When (6.67) is valid, we can treat $\mathbf{X}$ as the same as the previous sections when the sum and the product of such matrices are studied. We can study

$$\mathbf{X}_1 \cdots \mathbf{X}_L = \prod_{i=1}^{L} \mathbf{X}_i \tag{6.68}$$

where $\mathbf{X}_i, i = 1, \ldots, L$ are defined in (6.67). In practice, the asymptotic condition of $\rho \to \infty$ is approximately satisfied when ρ varies from 10 to 50. See Figure 6.8 and Figure 6.9 for illustration.

Theorem 6.5.1 Let the random matrix $\mathbf{X}$ be defined as in (6.66). Then, the empirical eigenvalue distribution of singular value equivalent of $\mathbf{X}$ converges almost surely to the limit

$$f_{\mathbf{H}_u}(z) = \begin{cases} \frac{1}{c\pi}\frac{1}{\sqrt{(1-\rho)^2+4\rho|z|^2}} & \sqrt{(1-\beta/\rho)(1-\beta)} \leqslant |z| \leqslant 1 \\ 0 & \text{elsewhere} \end{cases} \tag{6.69}$$

as $N, n, K \to \infty$ with the ratios $\rho = K/N$, and $c = N/n \leq 1$ fixed.

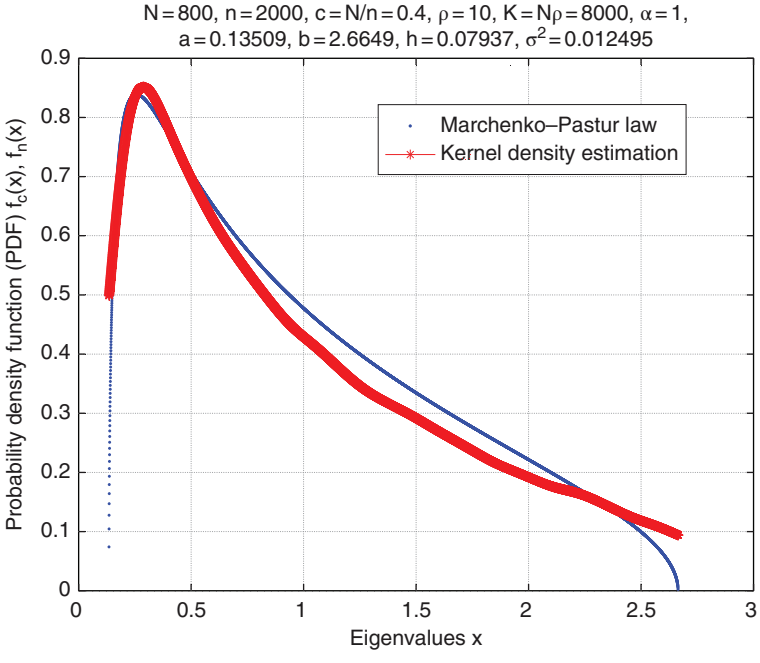


Figure 6.8 The empirical eigenvalue density function for one non-Hermitian random matrix: $N = 800, n = 2000, c = N/n = 0.4, a = 0.135, b = 2.66, h = 0.079$ and $\sigma^2 = 0.0124$ for $\rho = 10$ (thus $K = N\rho = 8000$).

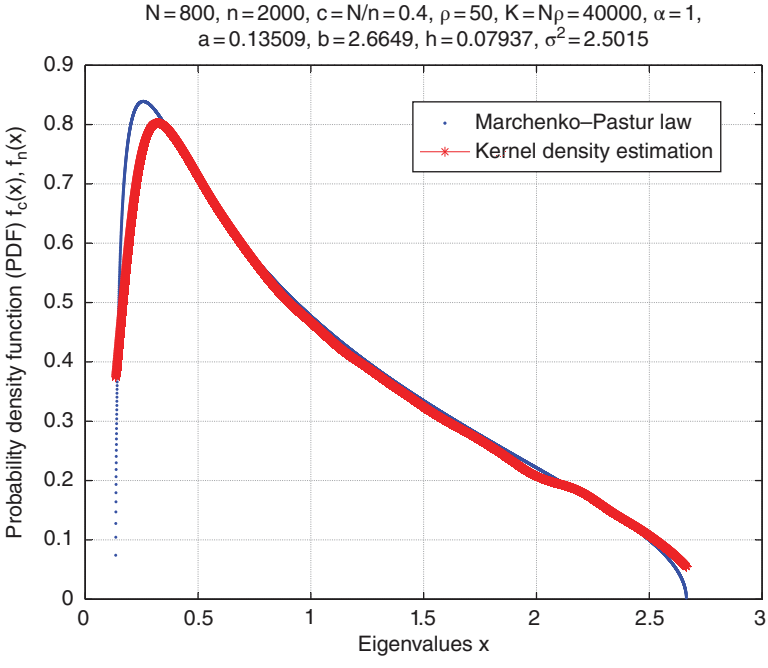


Figure 6.9 The same as Figure 6.8 except $\rho = 50$.

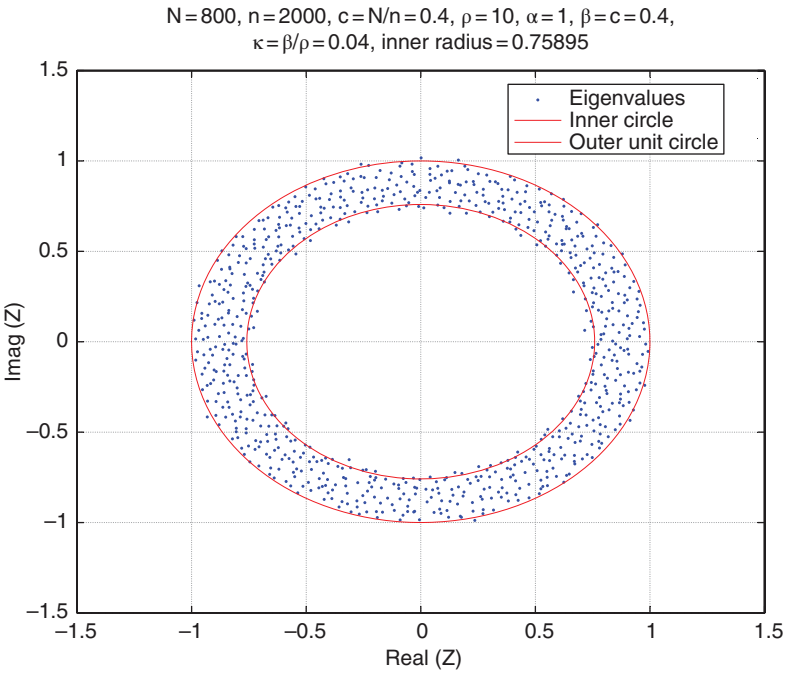


Figure 6.10 The eigenvalues for one non-Hermitian random matrix: $N = 800, n = 2000, c = N/n = 0.4, a = 0.135, b = 2.66, h = 0.079$ and $\sigma^2 = 0.0124$ for $\rho = 10$ (thus $K = N\rho = 8000$).

See Figure 6.10 and Figure 6.11 for the illustration of (6.69). The radius of the inner circle is given by $\sqrt{(1-\beta/\rho)(1-\beta)}$.

In Figure 6.12 and Figure 6.13, we study the case $L = 5$ in (6.67). The radius of the inner circle is given by $\left(\sqrt{(1-\beta/\rho)(1-\beta)}\right)^L$. For the general case, the author obtained it using a heuristic approach.

It is known in [270, 294] that free independence occurs asymptotically in large random matrices. As a result, when N and n go large, (6.67) can be viewed as the free multiplicative convolution of k free random variables. We are interested in the support of the k-time free multiplicative convolution of the measure μ with itself, which we denote as μ_k:

$$\mu_k = \underbrace{\mu \boxtimes \cdots \boxtimes \mu}_{k-\text{times}}$$

Let b_k denote the upper boundary of the support of μ_k:

Theorem 6.5.2 ([295]) Suppose that μ is a compactly supported probability measure on $\mathbb{R}^+$, with expectation 1 and variance σ^2. Then

$$\lim_{k\to\infty} \frac{b_k}{k} = e\sigma^2$$

where e denotes the base of natural logarithms, $e = 2.71\ldots$

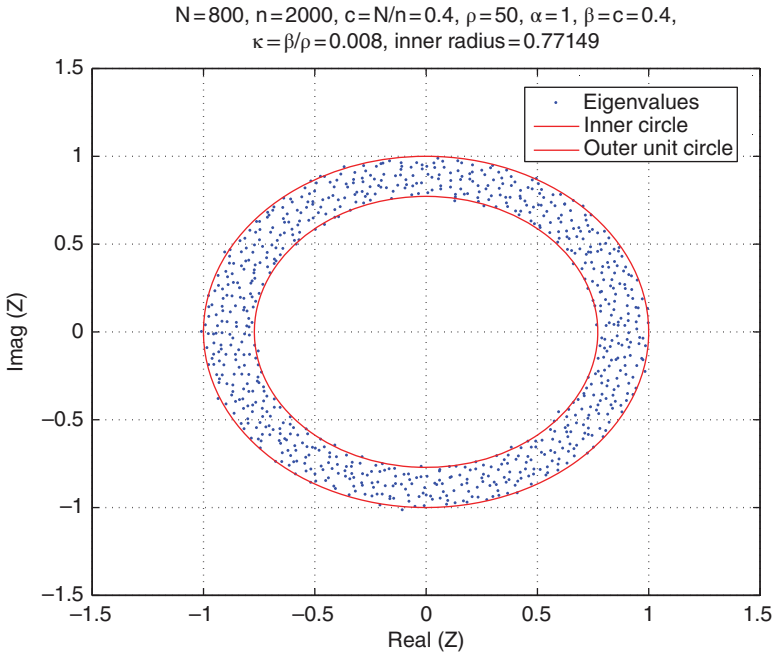


Figure 6.11 The same as Figure 6.10 except $\rho = 50$.

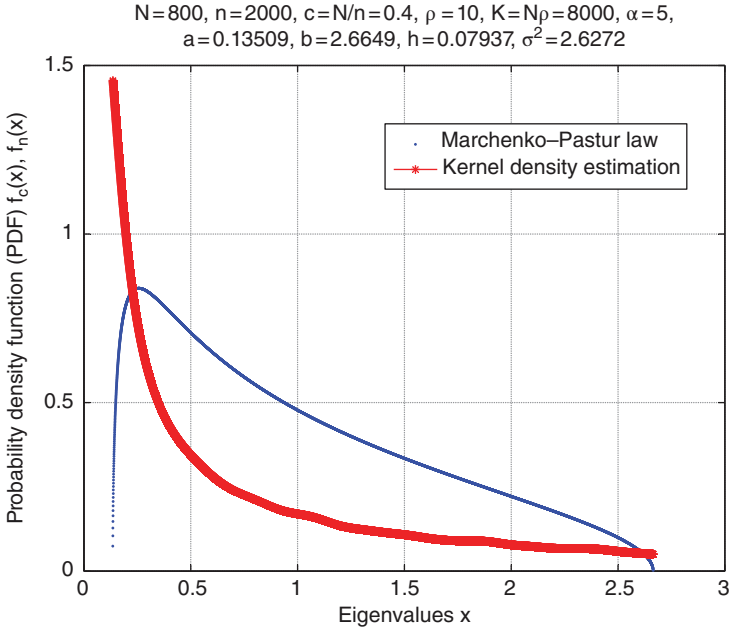


Figure 6.12 The eigenvalues for one non-Hermitian random matrix: $N = 800, n = 2000, c = N/n = 0.4, a = 0.135, b = 2.66, h = 0.079$ and $\sigma^2 = 0.0124$ for $\rho = 10$ and $L = 5$.

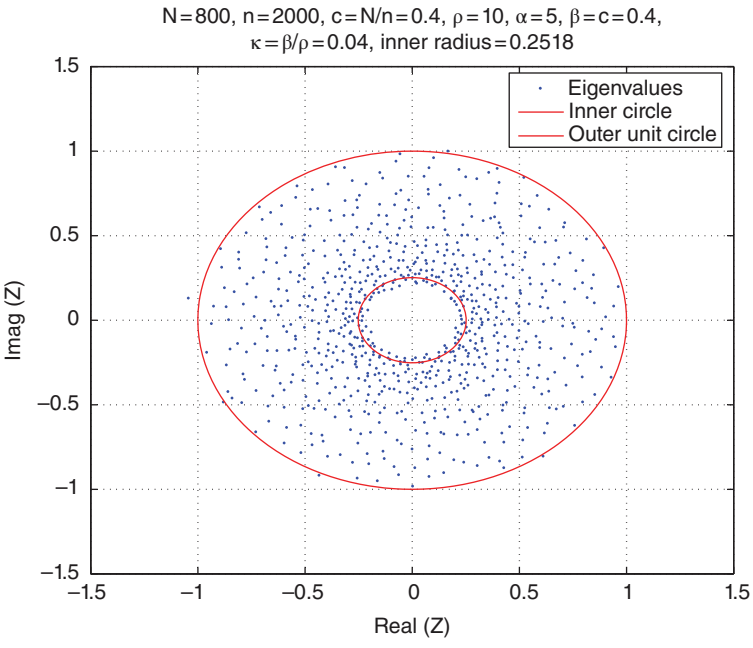


Figure 6.13 The same as Figure 6.12.

To say it very roughly, the largest singular value of the product approximately equals a certain average of all singular values multiplied by $\sqrt{n}$ and a proportionality constant.

From (3.10), we know the famous Marchenko–Pastur law is compacted-supported within the interval $[a_1, b_1]$, where a_1 and b_1 are defined in (6.56). To differentiate the "signal" from noise, we often need to know b_k, the upper boundary of the support of μ_k.

We can extend the above theorem.

Theorem 6.5.3 ((2012) [296]) There exists a universal constant $C > 0$ such that for all k and any $\mu_1, \ldots, \mu_k$ probability measures supported on $[0, b]$, satisfying $\mathbb{E}\left(\mu_i\right) = 0$ and $\mathrm{Var}\left(\mu_i\right) \geqslant \sigma^2$, for $i = 1, \ldots, k$, the the upper boundary B_k of the support of the measure $\mu_1 \boxtimes \cdots \boxtimes \mu_k$ satisfies

$$\sigma^2 k \leqslant B_k \leqslant Cbk.$$

In other words, for (not necessarily identically distributed) positive free random variables $\left(\mathbf{Y}_i\right)_{i\geqslant 1}$ such $\mathbb{E}\left(\mu_i\right) = 0$, $\mathrm{Var}\left(\mu_i\right) \geqslant \sigma^2$, and $\left\|\mathbf{Y}_i\right\| \leqslant b$, for $i \geq 1$, we have

$$\lim_{n\to\infty} \sup \frac{1}{K} \left\|\mathbf{Y}_1^{1/2} \cdots \mathbf{Y}_{K-1}^{1/2} \mathbf{Y}_K \mathbf{Y}_{K-1}^{1/2} \cdots \mathbf{Y}_1^{1/2}\right\| < Cb$$

and

$$\lim_{n\to\infty} \inf \frac{1}{K} \left\|\mathbf{Y}_1^{1/2} \cdots \mathbf{Y}_{K-1}^{1/2} \mathbf{Y}_K \mathbf{Y}_{K-1}^{1/2} \cdots \mathbf{Y}_1^{1/2}\right\| \geqslant \sigma^2$$

For compactly supported measures with mean 1, the variance is additive [296] with respect to free multiplicative convolution, that is

$$\mathrm{Var}\left(\mu_1 \boxtimes \cdots \boxtimes \mu_k\right) = \sum_{i=1}^{k} \mathrm{Var}\left(\mu_i\right)$$

MATLAB Code: Singular Value Equivalent Models

```
clear all;
%Reference
% Non-Hermitian Random Matrix Theory for MIMO Channels
% Burak Cakmak (2012) MS Thesis
% NTNU-Trondheim Norwegain University of
  Science and Technology
N=200*4; beta=0.4; rho=10; alpha=1; c=beta; n=N/c;
% c=N/n; c=p/n; beta=T/R=c.
R=N; T=n; kappa=beta/rho; S=R*rho; %kappa=beta/rho
radius_inner=((1-kappa)*(1-beta))^(alpha/2)

step=0.01/40; %step=0.01/10/4/2/2;
h=1/n^(1/3);
a=(1-sqrt(c))^2; b=(1+sqrt(c))^2;
x=(a+step):step:b;
fcx=(1/2/pi/c./x).*sqrt((b-x).*(x-a));
% Marcenko and Pastur law
H=bernoulli(0.5,N,N)+sqrt(-1)*bernoulli(0.5,N,N);
% i.i.d. complex matrix
U=H*sqrtm(inv(H'*H)); % Unitrary Haar matrix U of N x N
clear H;

Z=eye(N,N);
for i=1:alpha

H1=1/sqrt(2)*randn(S,T)+sqrt(-1)*1/sqrt(2)*randn(S,T);
H2=1/sqrt(2)*randn(R,S)+sqrt(-1)*1/sqrt(2)*randn(R,S);
Y=H2*H1/sqrt(R*T); % R x T (N x n)
Z=Z*U*sqrtm(Y*Y'); % singular value equivalent
end % i
clear H1; clear H2;clear Y; clear U;

VarZ=var(Z)';
sigma2=(mean(VarZ))^(1/alpha);

for j=1:N
Z(:,j)=Z(:,j)/std(Z(:,j)); % normalized the variance to one
end %j
```

```
Z=Z/sqrt(N); % normalized so the eigenvalues lie within
unit circle


lambda=eig(Z*Z'); % eigenvalues of sample covariance matrix


% kernel density estimation
L=(b-a)/step;
x1=a+step;
for j=1:L
 for i=1:N
   y=(x1-lambda(i))/h;
   Ky(i)=kernel(y);
 end %N
 fnx(j)=sum(Ky)/N/h;
 x1=x1+step;
 x2(j)=x1;
end %L


% figures
ifig=0;


ifig=ifig+1;figure(ifig)
hist(lambda);
xlabel('Eigenvalues x')
ylabel('Probability Density Function (PDF) f(x)')
 legend('Kernel Density Estimation');
title(['N=',int2str(N),', n=',int2str(n),',
c=N/n=',num2str(c),', a=',num2str(a),\ldots
 ', b=',num2str(b),', h=',num2str(h)])
grid


ifig=ifig+1;figure(ifig)
plot(x,n*h*(fcx-fnx))
xlabel('Eigenvalue x')
ylabel('Probability Density Function (PDF) ')
legend('Deviation n*h*[ f_c(x)-f_n(x) ]');
title(['N=',int2str(N),', n=',int2str(n),',
c=N/n=',num2str(c),',
a=',num2str(a),\ldots
 ', b=',num2str(b),', h=',num2str(h)])
grid
```

```
ifig=ifig+1;figure(ifig)
plot(x,fcx,'.b', x2,fnx,'-*r')
xlabel('Eigenvalues x')
ylabel('Probability Density
Function (PDF) f_c(x), f_n(x)')
 legend('Marcenko-Pastur Law','Kernel Density Estimation');
title(['N=',int2str(N),', n=',int2str(n),',
c=N/n=',num2str(c),\ldots
', \rho=',num2str(rho), ', K=N\rho=',num2str(S),\ldots ',
\alpha=',num2str(alpha),', a=',num2str(a),\ldots
 ', b=',num2str(b),', h=',num2str(h),',
\sigma^2=',num2str(sigma2)])
grid

ifig=ifig+1;figure(ifig);
lambdaZ=eig(Z);
t=0:2*pi/1000:2*pi;x=sin(t);y=cos(t); % unit circle
plot(real(lambdaZ),imag(lambdaZ),'.',
radius_inner*x,radius_inner*y,'r-',x,y,'r-');
 axis([-1.5 1.5 -1.5 1.5])
%xlabel('Eigenvalues x')
xlabel('real(Z)'); ylabel('imag(Z)');
legend('Eigenvalues','Inner Circle ','Outer Unit Circle');
title(['N=',int2str(N),',
n=',int2str(n),\ldots
 ', c=N/n=',num2str(c),', \rho=',num2str(rho),',
\alpha=',num2str(alpha),\ldots
 ', \beta=c=',num2str(beta), ',
\kappa=\beta/\rho=',num2str(kappa),\ldots
 ', inner radius=',num2str(radius_inner)])
grid

function [Kx] = kernel(x)
Kx=1/sqrt(2*pi)*exp(-0.5*x.^2);

function B=bernoulli(p,m,n);
% BERNOULLI.M
% This function generates n independent draws of a Bernoulli
% random variable with probability of success p.
% first, draw n uniform random variables

M = m;
N = n;
p = p;
B = rand(M,N) < p;
B=B*(-2)+ones(M,N);
```

6.6 The Power of the Non-Hermitian Random Matrix

The new results in this section were recently obtained by the author using a heuristic approach. This power of the non-Hermitian random matrix is reminiscent of the power mapping method that was used for noise reduction in the empirical covariance matrix in Section 2.10.5.

6.6.1 The Matrix Power

If $\mathbf{A}$ is diagonalizable, with $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^H$, where $\mathbf{U}$ is unitary, then for an arbitrary function we have the definition $g(\mathbf{A}) = \mathbf{U}g(\mathbf{D})\mathbf{U}^H$. Therefore for diagonalizable matrices [297, p.3], $g(\mathbf{A})$ has the same eigenvectors as $\mathbf{A}$ and its eigenvalues are obtained by applying g to those of $\mathbf{A}$. In particular, for $\alpha \in \mathbb{R}$, $\mathbf{A}^\alpha$ is a matrix function with $g(z) = z^\alpha = \exp(\alpha \ln z)$, where $z \in \mathbb{C}$. The definition of the power of a complex variable follows [298, Article 236]. For $z = |z| e^{j\phi} \in \mathbb{C}$, we have $z^\alpha = |z|^\alpha e^{j\alpha\phi}$. If α is a real number, the power α will not break the symmetry of z on the complex plane. For example, the circle $|z| = R$ is transformed into $|z|^\alpha = R^\alpha$. Consider a special case $\alpha = 1/M$, where M is a positive integer. Then

$$R^\alpha = \exp(\alpha \ln R) = \exp\left(\frac{1}{M} \ln R\right) \to 1 \tag{6.70}$$

as M goes large enough. Let $\lambda_i, i = 1, \ldots, N$ be the eigenvalues of $\mathbf{A}\mathbf{A}^H$, which are always positive. The eigenvalues of the matrix function $\left(\mathbf{A}\mathbf{A}^H\right)^\alpha$, for every $\alpha \in \mathbb{R}$, we have

$$\left(\lambda_i\right)^\alpha = \exp\left(\alpha \ln \lambda_i\right) \to 1, \quad \text{as } \alpha \to 0 \tag{6.71}$$

6.6.2 Spectrum

The eigenvalues of an $n \times n$ complex matrix $\mathbf{M}$ are the roots in $\mathbb{C}$ of its characteristic polynomial. We denote by $s_1(\mathbf{M}) \geqslant \ldots \geqslant s_n(\mathbf{M}) \geqslant 0$ the singular values of $\mathbf{M}$, defined for every $1 \leq i \leq n$ by the eigenvalues of the matrix $\sqrt{\mathbf{M}\mathbf{M}^H}$ $s_i(\mathbf{M}) = \lambda_i\left(\sqrt{\mathbf{M}\mathbf{M}^H}\right)$. We define the empirical spectral measure and the empirical singular values measure as

$$\mu_{\mathbf{M}} = \frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i(\mathbf{M})}, \quad \nu_{\mathbf{M}} = \frac{1}{n}\sum_{i=1}^{n} \delta_{s_i(\mathbf{M})}$$

$\mu_{\mathbf{M}}$ is a probability measure on $\mathbb{C}$ while and $\nu_{\mathbf{M}}$ is a probability measure on $\mathbb{C}^+$.

Let $\left(X_{ij}\right)_{i,j\geqslant 1}$ and $\left(Y_{ij}\right)_{i,j\geqslant 1}$ be independent, i.i.d. complex, random variables with mean 0 and variance 1. Similarly, let $\left(G_{ij}\right)_{i,j\geqslant 1}$ and $\left(H_{ij}\right)_{i,j\geqslant 1}$ be independent, complex, centered Gaussian variables with variance 1, independent of $\left(X_{ij}, Y_{ij}\right)$. We consider the random matrices

$$\mathbf{X}_n = \left(X_{ij}\right)_{1\leqslant i,j\leqslant n}, \mathbf{Y}_n = \left(Y_{ij}\right)_{1\leqslant i,j\leqslant n}, \mathbf{G}_n = \left(G_{ij}\right)_{1\leqslant i,j\leqslant n}, \text{ and } \mathbf{H}_n = \left(H_{ij}\right)_{1\leqslant i,j\leqslant n}$$

For ease of notation we will sometimes drop the subscript n. It is known that almost surely (a.s.) for n large enough, $\mathbf{X}$ is invertible and then $\mu_{\mathbf{X}^{-1}\mathbf{Y}}$ is a well defined random probability measure on $\mathbb{C}$. The generalized eigenvalues of $(\mathbf{A}, \mathbf{B})$ two $n \times n$ complex

matrices, are the zeros of the polynomial $\det(\mathbf{A} - z\mathbf{B})$. If $\mathbf{B}$ is invertible, it is simply the eigenvalues of $\mathbf{B}^{-1}\mathbf{A}$. Now, let μ be the probability measure whose density with respect to the Lebesgue measure on $\mathbb{C} \simeq \mathbb{R}^2$ is

$$\frac{1}{\pi\left(1+|z|^2\right)^2}$$

Through stereographic projection, μ is easily seen to be the uniform measure on the Riemann sphere.

Theorem 6.6.1 (Spherical ensemble [299–301]) For each integer $n \geq 1$,

$$\mathbb{E}\mu_{\mathbf{G}^{-1}\mathbf{H}} = \mu$$

We have a universality result.

Theorem 6.6.2 (Universality of generalized eigenvalues—Bordenave (2011) [302]) Almost surely,

$$\mu_{\mathbf{X}^{-1}\mathbf{Y}} - \mu_{\mathbf{G}^{-1}\mathbf{H}} \xrightarrow[n\to\infty]{} 0$$

Corollary 6.6.3 (Spherical law—Bordenave (2011) [302]) Almost surely

$$\mu_{\mathbf{X}^{-1}\mathbf{Y}} \xrightarrow[n\to\infty]{} \mu$$

Theorem 6.6.4 (Universality of sum and product of random matrices)—Bordenave (2011) [302]) For every integer n, let $\mathbf{M}_n, \mathbf{K}_n, \mathbf{L}_n$ be $n \times n$ complex matrices such that, for some positive real value $\alpha > 0$

- $x \mapsto x^{-\alpha}$ is uniformly bounded for $\left(\nu_{\mathbf{K}_n}\right)_{n\geqslant 1}$ and $\left(\nu_{\mathbf{L}_n}\right)_{n\geqslant 1}$, and $x \mapsto x^{\alpha}$ is uniformly bounded for $\left(\nu_{\mathbf{M}_n}\right)_{n\geqslant 1}$;
- for almost all $z \in \mathbb{C}$, $\nu_{\mathbf{K}_n^{-1}\mathbf{M}_n\mathbf{L}_n^{-1}-\mathbf{K}_n^{-1}\mathbf{L}_n^{-1}z}$ converges weakly to a probability measure ν_z.

Then, almost surely,

$$\mu_{\mathbf{M}+\mathbf{KXL}/\sqrt{n}} \xrightarrow[n\to\infty]{} \mu,$$

where μ depends only on $\left(\nu_z\right)_{z\in\mathbb{C}}$. For $\mathbf{M} = \mathbf{K} = \mathbf{L} = \mathbf{I}_n$, the $n \times n$ identity matrix, this statement gives the famous circular law theorem.

Lemma 6.6.5 (singular values of sum and product— [302]) If $\mathbf{A}, B$ are $n \times n$ complex matrices, for any real positive value $\alpha > 0$

$$\int x^{\alpha} d\nu_{\mathbf{A}+\mathbf{B}}(x) \leqslant 2^{1+\alpha}\left(\int x^{\alpha} d\nu_{\mathbf{A}}(x) + \int x^{\alpha} d\nu_{\mathbf{B}}(x)\right)$$

$$\int x^{\alpha} d\nu_{\mathbf{AB}}(x) \leqslant 2\left(\int x^{\alpha} d\nu_{\mathbf{A}}(x)\right)^{1/2}\left(\int x^{\alpha} d\nu_{\mathbf{B}}(x)\right)^{1/2}$$

6.6.3 The Product

Given $\tilde{\mathbf{X}}_i, i = 1, \ldots, L$ in $\mathbb{C}^{N\times n}$, we can always first obtain the singular value equivalent

$$\mathbf{X}_i = \mathbf{U}\sqrt{\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i^H}$$

where $\mathbf{U}$ is an $N \times N$ Haar unitary matrix. For two arbitrary positive integers L and M, we study the product of non-Hermitian random matrices followed by a power $1/M$

$$\left(\mathbf{X}_1 \cdots \mathbf{X}_L\right)^{1/M} \tag{6.72}$$

or

$$\left(\mathbf{X}_1\right)^{1/M} \cdots \left(\mathbf{X}_L\right)^{1/M} \tag{6.73}$$

where $\mathbf{X}_i, i = 1, \ldots, L$ are $N \times N$ non-Hermitian random matrices, whose entries are i.i.d. with mean zero and variance $1/N$. (6.72) and (6.73) will be shown to be identical. This is the basis for defining

$$\mathbf{X}^{L/M} = \left(\mathbf{X}_1 \cdots \mathbf{X}_L\right)^{1/M} = \left(\mathbf{X}_1\right)^{1/M} \cdots \left(\mathbf{X}_L\right)^{1/M}$$

Now it is well known that the asymptotic eigenvalue distribution of the product $\mathbf{X}_1 \cdots \mathbf{X}_L$ is identical to $(\mathbf{X}_i)^L, i = 1, \ldots, L$, according to Theorem 6.2.5. We rewrite (6.72) as $\mathbf{X}^L$ to represent any one of the L matrices. Then we obtain $\mathbf{X}^{L/M}$, which is natural. When $M = L$, (6.72) is the geometric mean of the L matrices. Thus (6.72) reduces to $\mathbf{X}$. This result is intuitive. The geometric mean of an arbitrary number of large non-Hermitian random matrices is identical to any one such large matrix.

What is unexpected is the discovery of the results when M is significantly larger than L. Define the ratio $\alpha = L/M$. When α is smaller than $1/5$, the asymptotic eigenvalue distribution is found to follow Gaussian distribution (Figure 6.14) with the mean equal to one. All the eigenvalues are distributed very close to the unit circle, as illustrated in Figure 6.15. What matters is the ratio α. The result looks the same if we use $L = 30$ and $M = 150$.

In analogy with a complex number z, the order of L and M will not affect the result. In other words, $\left(\mathbf{X}^L\right)^{1/M}$ and $\left(\mathbf{X}^{1/M}\right)^L$ are identical, as illustrated in Figure 6.16 and Figure 6.17. Therefore, we have the unambiguous definition for the fractional power α, i.e., $\mathbf{X}^\alpha$. The radius of the inner circle on the complex plane is found to be

$$r_{\text{in}} = \left(\sqrt{1-c}\right)^{(L/M)^2} \tag{6.74}$$

where $c = N/n$. Using (6.70), we can always choose L and M to force the inner circle to be close enough to the outer unit circle. The empirical expression (6.74) has been tested for a large class of parameters such that $|z| \geq r_{\text{in}}$. Using the rescaling in the MATLAB code, we always guarantee that the upper bound of the spectrum is the unit circle $|z| = 1$.

```
clear all;
L=3;M=15; N=200*2; beta=0.3; c=beta; n=N/c; % c=N/n;
c=p/n; beta=T/R=c; beta<1.
step=0.01/40; %step=0.01/10/4/2/2;
h=1/n^(1/3);
a=(1-sqrt(c))^2; b=(1+sqrt(c))^2;
```

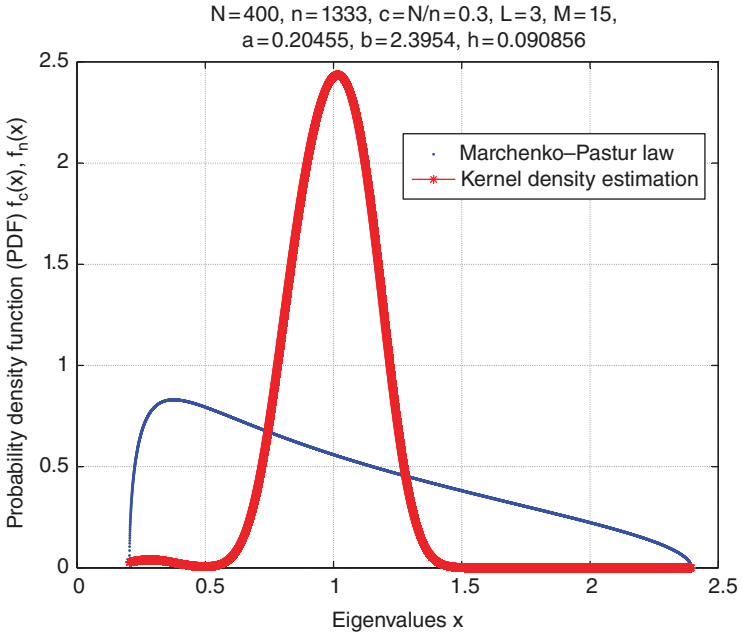


Figure 6.14 The eigenvalues of $\left(\mathbf{X}^{L}\right)^{1/M}$ for one non-Hermitian random matrix $\mathbf{X}$ of size $N \times n$: $N = 400, n = 1333, c = N/n = 0.3, L = 3, M = 15, a = 0.135, b = 2.66, h = 0.0908$. The ratio α is $\alpha = L/M = 1/5$.

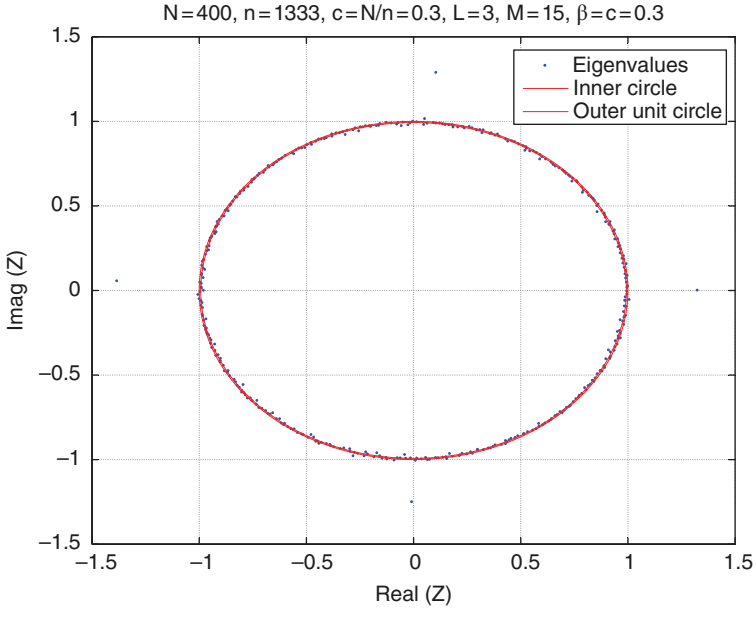


Figure 6.15 The same as Figure 6.14. Four outliers.

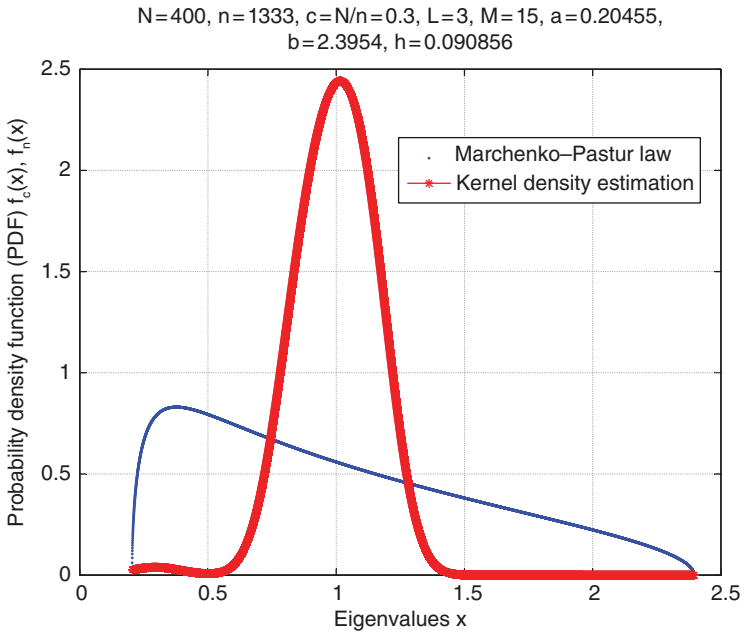


Figure 6.16 The eigenvalues of $\left(\mathbf{X}^{1/M}\right)^{L}$ for one non-Hermitian random matrix $\mathbf{X}$ of size $N \times n$. All other parameters are the same as Figure 6.14.

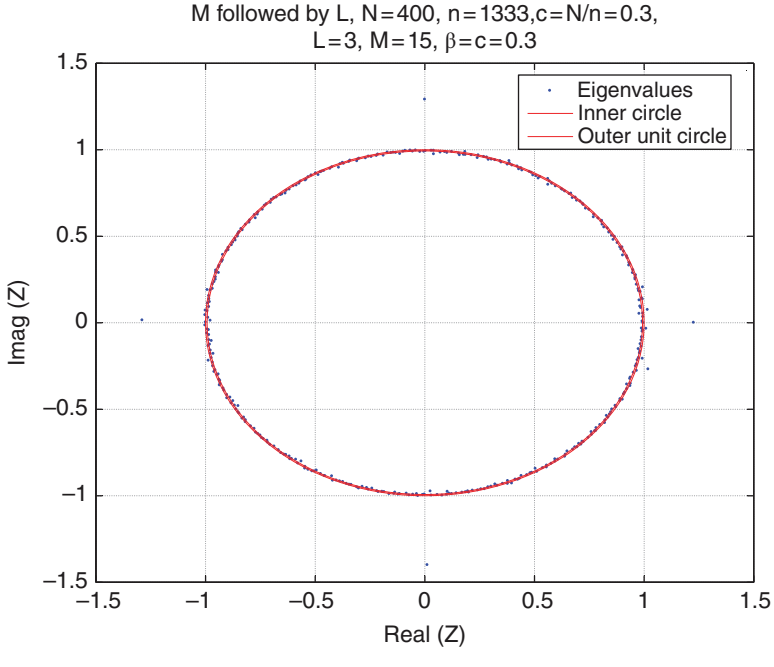


Figure 6.17 The same as Figure 6.16. Four outliers.

```
x=(a+step):step:b;
fcx=(1/2/pi/c./x).*sqrt((b-x).*(x-a)); % Marcenko
and Pastur law
X=bernoulli(0.5,N,N)+sqrt(-1)*bernoulli(0.5,N,N);
% i.i.d. complex matrix X
U=X*sqrtm(inv(X'*X)); % Unitrary Haar matrix U of N x N
 %X=1/sqrt(2)*randn(N,n)+sqrt(-1)*1/sqrt(2)*randn(N,n);
 % Gaussian random matrix
X=1/sqrt(2)*bernoulli(0.5,N,n)+sqrt(-1)*1/sqrt(2)
*bernoulli(0.5,N,n);
% Bernoulli random matrix
D=zeros(N,N); D(1,1)=1.3; D(2,2)=j*1.3;
D(3,3)=-j*1.3; D(4,4)=-1.3;
X=U*sqrtm(X*X'); % singular value equivalent
X=U*(X*X')^(L/2); % singular value equivalent X^(L)
X=U*(X*X')^(1/2/M); % singular value equivalent X^(1/M)
radius_inner=(1-beta)^(1/2*(L/M)^2) % Y=X^L; Z=Y^(1/M);
Z=X;
for j=1:N
Z(:,j)=Z(:,j)/std(Z(:,j)); % normalized the variance to one
end %j
A=U*D*U'*sqrt(N); Z=Z+A; % A will cause outliers
Z=Z/sqrt(N); % normalized so the eigenvalues lie within
unit circle
lambda=eig(Z*Z'); % eigenvalues of sample covariance matrix
% kernel density estimation
Mtemp=(b-a)/step;
x1=a+step;
for j=1:Mtemp
 for i=1:N
   y=(x1-lambda(i))/h;
   Ky(i)=kernel(y);
 end %N
 fnx(j)=sum(Ky)/N/h;
 x1=x1+step;
 x2(j)=x1;
end %L
```

To save space, other functions are omitted; they can found in other codes that were included previously.

6.7 Power Series of Large Non-Hermitian Random Matrices

The notation follows Section 6.6. In Section 6.6 we precisely defined the fractional power $\mathbf{X}^{\alpha}$ for a square matrix $\mathbf{X}$ and $\alpha \in \mathbb{R}$, in particular $\alpha = L/M$ where L and

M are arbitrary positive integers. With suitable rescaling, the spectrum of $\mathbf{X}^{\alpha}$ lies within a single ring with the inner circle and the outer unit circle. The radius of the inner circle is given by (6.74). For a complex number z, the power z^{α} will not break the symmetry of z on the complex plane. In this section, the functions $f(z)$ will generally not keep this symmetry. Even in this context, we will find that it is very beneficial to pro-process the data matrix $\mathbf{X}$ with the power function $\mathbf{X}^{\alpha}$, which, with properly chosen α, transforms the all the eigenvalues of $\mathbf{X}$ to possess special properties: (1) Gaussian distribution; (2) mean one. By removing the mean and normalizing the variance, the resultant eigenvalues of $\mathbf{X}^{\alpha}$ are Gaussian with zero-mean and variance one.

Now we are in a position to study the general power series

$$\sum_{k=0}^{\infty} a_k \mathbf{X}^k$$

where it is assumed that the coefficients a_k satisfy certain conditions for the power series to exist. We are motivated by the fact that the power $\mathbf{X}^k$ is a R-diagonal matrix, so $\mathbf{X}^k$ has a complete analogy with complex number $z = |z|^k e^{jk\phi}$, where $z = |z|\, e^{j\phi}$ is a complex number. One naturally thinks of the power series [298, 303, 304]

$$\sum_{k=0}^{\infty} b_k z^k.$$

We can replace the complex argument z with $\mathbf{X} = UP$, where $\mathbf{P} = \sqrt{\mathbf{X}\mathbf{X}^H}$ is the polar part of the matrix $\mathbf{X}$, and $\mathbf{U}$ is the Haar unitary matrix.

6.7.1 The Geometric Series

The geometric series is defined as the series

$$1 + z + z^2 + z^3 + z^4 + \cdots$$

Consider the series of moduli

$$1 + |z| + |z|^2 + |z|^3 + |z|^4 + \cdots$$

for this series

$$\begin{aligned} S_{n,p} &= |z|^{k+1} + |z|^{k+2} + \cdots + |z|^{k+p} \\ &= |z|^{k+1} \tfrac{1-|z|^p}{1-|z|} \end{aligned}$$

If $|z| < 1$, then $S_{n,p} < |z|^{k+1} \frac{1}{1-|z|}$ for all values of p. The series

$$1 + |z| + |z|^2 + |z|^3 + |z|^4 + \cdots$$

is convergent so long as $|z| < 1$, and therefore the geometric series is absolutely convergent if $|z| < 1$. When $|z| \geq 1$, the terms of the geometric series do not tend to zero as k tends to infinity, and the series is therefore convergent.

When $g(z)$ is a polynomial or rational function [297, pp.1–2] with scalar coefficients and a scalar argument, z, it is natural to define $g(\mathbf{A})$ by replacing the complex argument z with $\mathbf{A}$, replacing division by matrix inversion, and replacing 1 with the identity matrix.

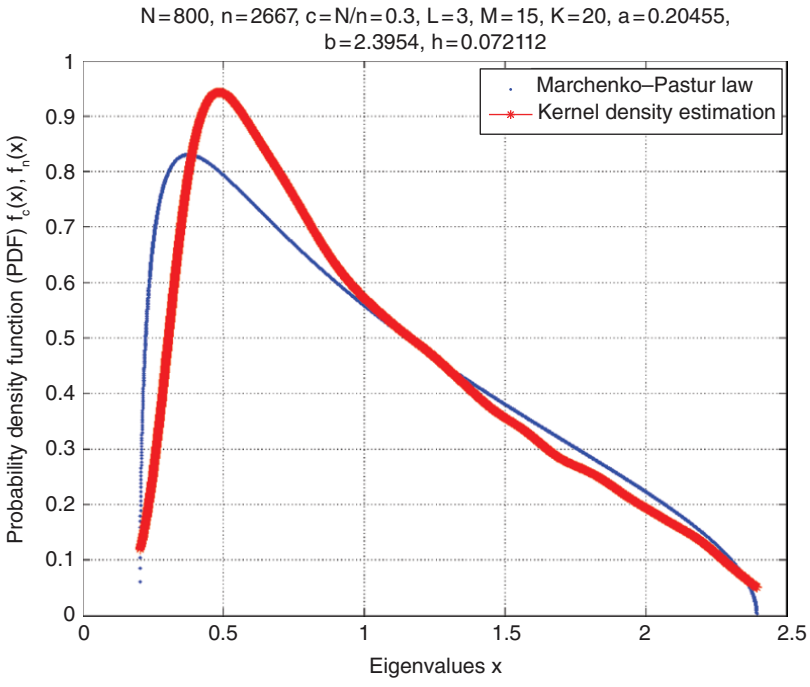


Figure 6.18 The eigenvalues of a geometric series of K terms: each term is $\left(\mathbf{X}^{L/M}\right)$ for one non-Hermitian random matrix $\mathbf{X}$ of size $N \times n$. $N = 800, n = 2667, c = 0.3, L = 3, M = 15, K = 20, a = 0.204, b = 2.395$.

The spectrum is upper bounded by the unit circle, so the geometric series is convergent if the z is replaced by $\mathbf{X}^{L/M}/(1+\varepsilon), \varepsilon > 0$. The parameter ε is chosen to guarantee the convergence of the geometric series. In practice, we are interested in a finite sum

$$f(\mathbf{Y}) = \mathbf{I} + \mathbf{Y} + \cdots + \mathbf{Y}_K = \sum_{k=1}^{K} \mathbf{Y}^k, \qquad \mathbf{Y} = \mathbf{X}^{L/M}$$

The radius of the inner circle for $f(\mathbf{Y})$ is found to be

$$r_{\text{in}} = r^K, \qquad r = \left(\sqrt{1-c}\right)^{(L/M)^2} \tag{6.75}$$

where $c = N/n$. The outer boundary of the spectrum is the unit circle. Comparing (6.75) and (6.75), the integral term by term affects the radius of the inner circle.

In Figure 6.18 and Figure 6.19, we consider the geometric series with $K = 20$. We can always choose a bigger ratio L/M such that the inner circle is very close to the unit circle.

6.7.2 Power Series

A power series may or may not converge for points that are actually on the periphery of the circle; thus the series

$$1 + \frac{z}{1^s} + \frac{z^2}{2^s} + \frac{z^3}{3^s} + \frac{z^4}{4^s} + \cdots \tag{6.76}$$

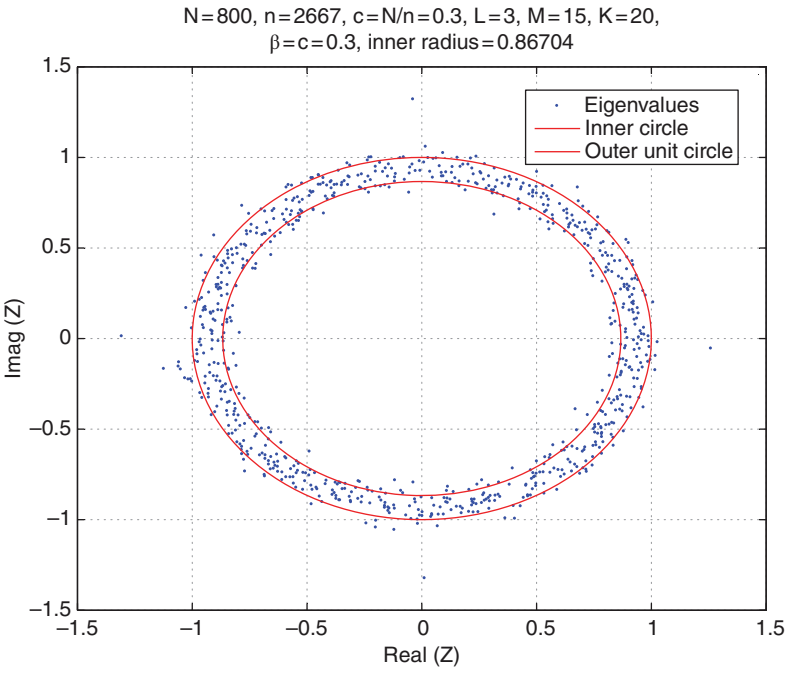


Figure 6.19 The same as Figure 6.18. Four outliers.

whose radius of convergence is unity, converges or diverges at the point $z = 1$ according as s is greater or not greater than unity.

Let

$$a_0 + a_1 z + a_2 z^2 + a_3 z^3 + a_4 z^4 + \cdots$$

be a power series and consider the series

$$a_1 + a_2 2z + 3a_3 z^2 + 4a_4 z^3 + \cdots$$

which is obtained by differentiating the power series term by term. The derived series has the same circle of convergence as the original series. The series

$$\sum_{k=0}^{\infty} a_k \frac{z^{k+1}}{k+1} \tag{6.77}$$

obtained by integrating the original power series term by term, has the same circle of convergence as $\sum_{k=0}^{\infty} a_k z^{k+1}$. For (6.76), we have $a_0 = 1, a_k = \frac{1}{k^s}, k \geq 1$, thus we obtain the function

$$f(z) = z + \sum_{k=1}^{\infty} \frac{1}{k^s} \frac{1}{k+1} z^{k+1} = z + \frac{1}{1^s} \frac{z^2}{2} + \frac{1}{2^s} \frac{z^3}{3} + \frac{1}{3^s} \frac{z^4}{4} + \cdots \tag{6.78}$$

The series

$$1 + \frac{z}{1!} + \frac{z^2}{2!} + \cdots + \frac{z^k}{k!} + \cdots$$

is a power series converging *everywhere*, and thus define a function regular in the whole plane. To every point z of the complex plane, there corresponds a definite number w, the sum of the above series. This function may be used to define powers of the base e for all complex powers $\mathbf{z}$, called the exponential function $e^z = \exp(z)$. If p is a positive number, we define

$$p^z = \exp(z \ln p), \quad z^p = \exp(p \ln z)$$

where ln is the natural logarithm.

Now we can replace the complex argument z with the $\mathbf{Y}$ in (6.78), to obtain

$$f(\mathbf{Y}) = \mathbf{Y} + \sum_{k=1}^{\infty} \frac{1}{k^s} \frac{1}{k+1} \mathbf{Y}^{k+1} = \mathbf{Y} + \frac{1}{1^s} \frac{\mathbf{Y}^2}{2} + \frac{1}{2^s} \frac{\mathbf{Y}^3}{3} + \frac{1}{3^s} \frac{\mathbf{Y}^4}{4} + \cdots \tag{6.79}$$

For instance, we can set $\mathbf{Y} = \mathbf{X}^{L/M}$. Similarly, we can do this for the geometric series (6.76).

We consider the series (6.77) where a_k is the coefficients for the geometric series. The radius of the inner circle in this case is found to be

$$r_{\text{in}} = r^{K+1}, \qquad r = \left(\sqrt{1-c}\right)^{(L/M)^2} \tag{6.80}$$

where $c = N/n$. The outer boundary of the spectrum is the unit circle. Simulations agree very closely with (6.80).

Using the geometric series (6.76), we replace the argument with $\mathbf{Z}$ defined as

$$\mathbf{Z} = SNR \cdot \mathbf{I}_N + \mathbf{X} \tag{6.81}$$

where SNR is the signal-to-noise ratio $SNR = \text{Tr}\left(\mathbf{I}_N\right) / \text{Tr}\left(\mathbf{X}\mathbf{X}^H\right)$. Here the entries of $\mathbf{X}$ are i.i.d. with zero mean and variance one. In Figures 6.20 and 6.21 we can visualize the difference of the perturbation caused by the signal term for different SNRs. Rich mathematical structures are observed, as functions of K and M. The complex $s = |s| \exp(\text{Angle})$ is used in the geometric series (6.76). The phase angle of s plays a significant role in this data visualization.

Now we summarize our algorithm as follows:

1) Given L rectangular complex matrices $\tilde{\mathbf{X}}_i, i = 1, \ldots, L$, we find the singular value equivalent $\mathbf{X}_i, i = 1, \ldots, L$, which are square matrices.
2) Take the power function of the matrix, $\mathbf{X}_i^{\alpha}$ for each $\mathbf{X}_i$. In particular, $\alpha = L/M$.
3) We do the same steps as above even when there is additive signal.
4) Use the symmetry-breaking series such as geometric series to form plots in the complex plane to guide us for better visualization.

□

We are naturally motivated to study

$$\begin{aligned} \mathcal{H}_0 &: \quad \mathbf{Z} = \mathbf{X}^{\alpha} \\ \mathcal{H}_1 &: \quad \mathbf{Z} = \left(SNR \cdot \mathbf{I}_N + \mathbf{X}\right)^{\alpha}, \quad \alpha \in \mathbb{R} \end{aligned} \tag{6.82}$$

For any complex exponent α and any complex z, the binomial series

$$(1+z)^{\alpha} = \sum_{k=0}^{\infty} \binom{\alpha}{k} z^k = 1 + \binom{\alpha}{1} z + \binom{\alpha}{2} z^2 + \cdots + \binom{\alpha}{k} z^k + \cdots$$

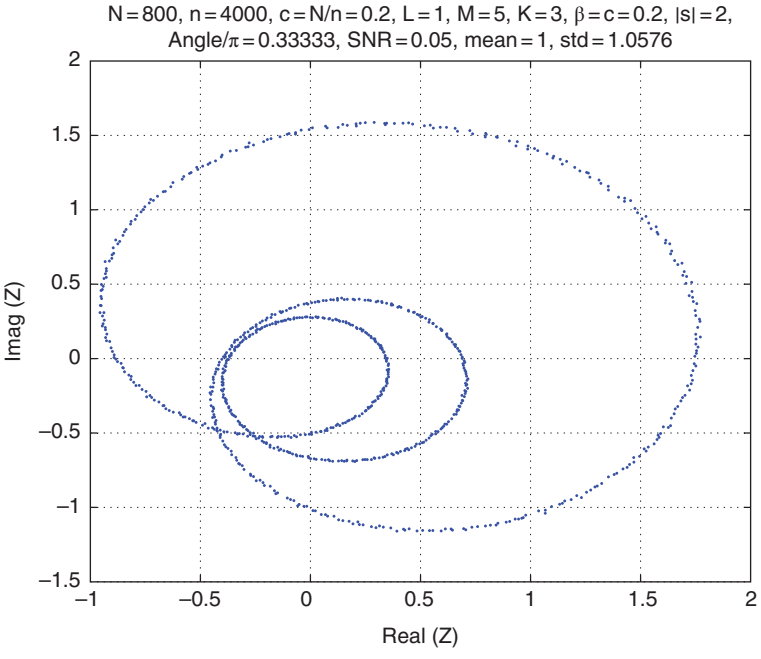


Figure 6.20 The eigenvalues of a geometric series of K terms: each term is $\left(\mathbf{X}^{L/M}\right)$ for one non-Hermitian random matrix $\mathbf{X}$ of size $N \times n$. $N = 800$, $n = 2667$, $c = 0.3$, $L = 1$, $M = 5$, $K = 3$, $|s| = 2$, $\text{Angle} = \pi/3$, and $SNR = 0.05$.

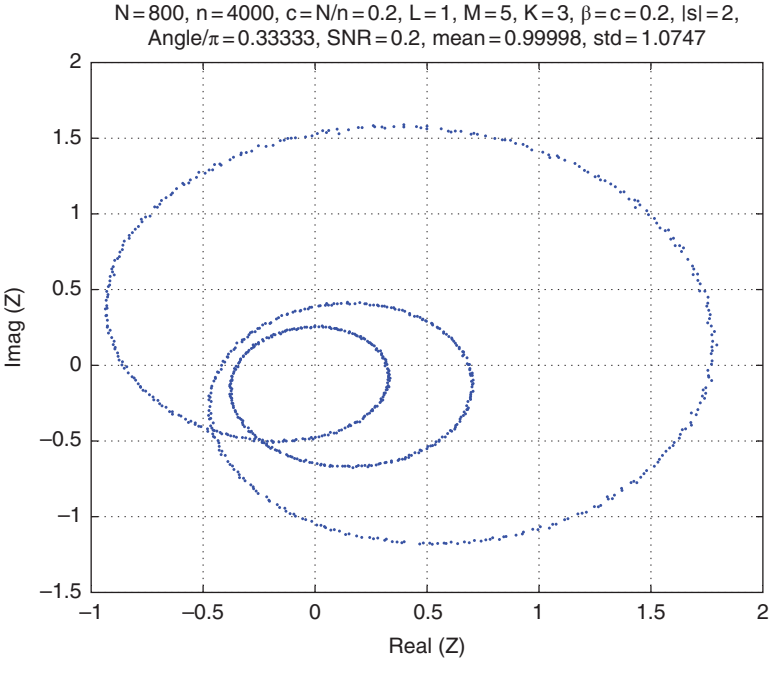


Figure 6.21 The same as Figure 6.20 except $SNR = 0.2$.

converges and has for sum the principal value of the power $(1+z)^{\alpha}$. We define the symbol

$$\binom{\alpha}{0} = 1, \binom{\alpha}{k} = \frac{\alpha(\alpha-1)\cdots(\alpha-k+1)}{1\cdot 2\cdot \quad \cdots \quad k}, \text{for } k \geq 1$$

and for every real α.

```
clear all;
L=1;M=5; K=3; N=200*4; beta=0.2; c=beta; n=N/c;
% c=N/n; c=p/n; beta=T/R=c; beta<1.
d=2; angle=pi/3; ifig=0; s=d*(cos(angle)
+sqrt(-1)*sin(angle));
X=bernoulli(0.5,N,N)+sqrt(-1)*bernoulli(0.5,N,N);
% i.i.d. complex matrix X
U=X*sqrtm(inv(X'*X)); % Unitrary Haar matrix U of N x N
 %X=1/sqrt(2)*randn(N,n)+sqrt(-1)*1/sqrt(2)*randn(N,n);
 % Gaussian random matrix
X=1/sqrt(2)*bernoulli(0.5,N,n)+sqrt(-1)*1/sqrt(2)
*bernoulli(0.5,N,n);
% Bernoulli random matrix



X=U*sqrtm(X*X'); % singular value equivalent
 S=eye(N,N)*sqrt(N); X=SNR*S+X;
X=U*(X*X')^(L/2); % singular value equivalent X^(L)
X=U*(X*X')^(1/2/M); % singular value equivalent X^(1/M)
X=geometricseries(X,K,s);
 radius_inner=(1-beta)^(1/2*(L/M)^2); % Y=X^L; Z=Y^(1/M);
Z=X;
for j=1:N
Z(:,j)=Z(:,j)/std(Z(:,j)); % normalized the variance to one
end %j
Z=Z/sqrt(N); % normalized so the eigenvalues lie within
unit circle
lambda=eig(Z*Z'); % eigenvalues of sample covariance matrix
lambdaZ=eig(Z);
ifig=kerneldensity(lambda,c,L,M,K,ifig);


function ifig=kerneldensity(lambda,c,L,M,K,ifig)


N=length(lambda); n=round(N/c); step=0.001+0.01/40;
%step=0.01/10/4/2/2;
h=1/n^(0.33);
```

```
a=(1-sqrt(c))^2; b=(1+sqrt(c))^2; x=(step):step:3;
fcx=(1/2/pi/c./x).*sqrt((b-x).*(x-a)); % Marcenko
and Pastur law


% kernel density estimation
x1=step; Mtemp=(3-x1)/step;
for j=1:Mtemp
 for i=1:N
 y=(x1-lambda(i))/h; Ky(i)=kernel(y);
 end %N
 fnx(j)=sum(Ky)/N/h; x1=x1+step; x2(j)=x1;
end %L
```

Missing functions can be found in the previous codes.

6.8 Products of Random Ginibre Matrices

For any integer number k, there exists a probability measure $\pi(k)$, called the Fuss–Catalan distribution of order k, whose moments are the generalized Fuss–Catalan numbers given in terms of the binomial symbol

$$\int_0^{b(k)} x^m \pi^{(k)}(x)\,dx = \frac{1}{km+1}\binom{km+m}{m} =: FC_m^{(k)} \tag{6.83}$$

The measure $\pi^{(k)}(x)$ has no atoms (or Dirac measures), and it is supported on $[0, b(k)]$ where $b(k) = (k+1)^{k+1}/k^k$. Its density is analytic on $[0, b(k)]$, and bounded at $x = b(k)$, with asymptotic behavior $\sim 1/\left(\pi x^{k/(k+1)}\right)$ at $x \to 0$. This distribution arises in random matrix theory as one studies the product of k independent random square Ginibre matrices, $\mathbf{Z} = \prod_{i=1}^{k} \mathbf{G}_i$. In this case the squared singular values of $\mathbf{Z}$, i.e., the eigenvalues of $\mathbf{Z}\mathbf{Z}^H$ have asymptotic distribution $\pi^{(k)}(x)$ in the large matrix limit. The same Fuss–Catalan distribution also describes asymptotically the statistics of singular values of k-th power of a single random Ginibre matrix [305]. In terms of free probability theory, it is the free multiplicative convolution product of k copies of the Marchenko–Pastur distribution [306, 307], which is written as $\pi^{(k)}(x) = \left[\pi^{(1)}\right]^{\boxtimes k}$. $\boxplus$ and $\boxtimes$ represent, respectively, the free additive convolution and the free multiplicative convolution. They are Voiculescu's operations $\boxplus$ and $\boxtimes$.

An explicit expression of the spectral density for $k = 2$ is given by

$$\pi^{(2)}(x) = \frac{2^{1/3}\sqrt{3}}{12\pi} \frac{\left[2^{1/3}\left(27 + 3\sqrt{81 - 12x}\right)^{2/3} - 6x^{1/3}\right]}{x^{2/3}\left(27 + 3\sqrt{81 - 12x}\right)^{1/3}} \tag{6.84}$$

where $x \in [0, 27/4]$. See Figure 6.22.

Making use of the inverse Mellin transform and the Meijer G-function, one may find a more explicit form of this distribution, as a superposition of hyper-geometric functions of the type ${}_kF_{k-1}$

$$\pi^{(k)}(x) = \sum_{i=1}^{k} \Lambda_{i,k} x^{\frac{i}{k+1}-1} {}_kF_{k-1}\left(\left[\{a_j\}_{j=1}^{k}\right]; \left[\{b_j\}_{j=1}^{k}, \{b_j\}_{j=i+1}^{k}\right]; \frac{k^k}{(k+1)^{k+1}}x\right) \tag{6.85}$$

where

$$a_j = 1 - \frac{1+j}{k} + \frac{i}{k+1}, \quad b_j = 1 + \frac{i-j}{k+1}$$

and the coefficients $\Lambda_{i,k}$ read for $i = 1, 2, \ldots, k$

$$\Lambda_{i,k} = \frac{1}{k^{3/2}} \sqrt{\frac{k+1}{2\pi}} \left(\frac{k^{k/(k+1)}}{k+1}\right)^i \frac{\left[\prod_{j=1}^{i-1} \Gamma\left(\frac{j-i}{k+1}\right)\right] \left[\prod_{j=i+1}^{k} \Gamma\left(\frac{j-i}{k+1}\right)\right]}{\prod_{j=1}^{k} \Gamma\left(\frac{j+1}{k} - \frac{n}{k+1}\right)} \tag{6.86}$$

Here ${}_pF_q\left(\left[\{a_j\}_{j=1}^{p}\right]; \left[\{b_j\}_{j=1}^{q}\right]; x\right)$ stands for the hypergeometric function [308] of the type ${}_pF_q$ with p "upper" parameters a_j and q "lower" parameters b_j of the real argument x.The symbol $\{a_j\}_{j=1}^{r}$ represents the list of r elements, $a_1, a_2, \ldots, a_r$. The above distribution is exact and it describes the density of squared singular values of k square Ginibre matrices in the limit of large matrix size N. Observe that, in the simplest case $k = 1$, the above form reduces to the Marchenko–Pastur distribution,

$$\pi^{(1)}(x) = \frac{1}{\pi\sqrt{x}} {}_1F_0\left(\left[-\frac{1}{2}\right]; []; \frac{1}{4}x\right) = \frac{\sqrt{1-x/4}}{\pi\sqrt{x}} \tag{6.87}$$

while the case $k = 2$

$$\pi^{(2)}(x) = \frac{\sqrt{3}}{2\pi x^{2/3}} {}_2F_1\left(\left[-\frac{1}{6}, \frac{1}{3}\right]; \left[\frac{2}{3}\right]; \frac{4}{27}x\right) - \frac{\sqrt{3}}{6\pi x^{1/3}} {}_2F_1\left(\left[\frac{1}{6}, \frac{2}{3}\right]; \left[\frac{4}{3}\right]; \frac{4}{27}x\right) \tag{6.88}$$

is equivalent to the form (6.84). See Figure 6.22 and Figure 6.23 for the illustration of the product of $k = 2$ matrices.

The upper edge $b(k) = (k+1)^{k+1}/k^k$ of the Fuss–Catalan distribution $\pi^{(k)}(x)$ for large matrices determines the size of the largest eigenvalue λ_{max} of the sample covariance matrix of size N. For $k = 1$, we obtain $b(1) = 4$ so that $\lambda_{max} \approx 4/N$.

Consider the following ensemble of non-Hermitian random matrices parametrized by an arbitrary m–dimensional probability vector $\mathbf{w} = w_1, \ldots, w_m$ and a non-negative integer k:

$$\mathbf{Z} := \left[w_1\mathbf{U}_1 + w_2\mathbf{U}_2 + \cdots + w_m\mathbf{U}_m\right] \mathbf{G}_1 \cdots \mathbf{G}_k \tag{6.89}$$

Here $\mathbf{U}_1, \ldots, \mathbf{U}_m$ of $N \times N$ denote m independent random unitary matrix distributed according to the Haar measure, while $\mathbf{G}_1, \ldots, \mathbf{G}_k$ are independent square random

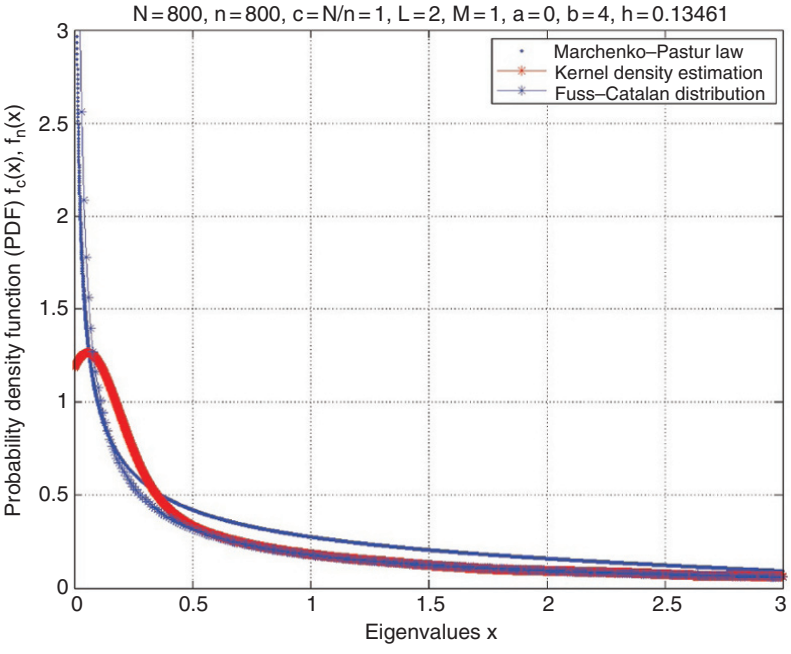


Figure 6.22 Product of $k = 2$ square i.i.d. matrices. $N = 800$, $n = 100$, $c = N/n = 1$.

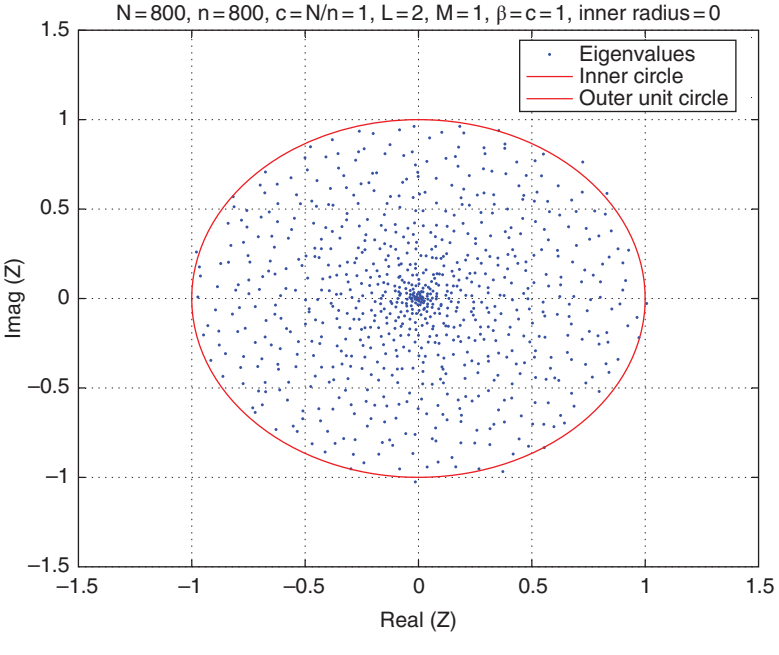


Figure 6.23 The same as Figure 6.22.

matrices of size N from the complex Ginibre ensemble. The empirical sample covariance matrix is obtained as a normalized Wishart-like matrix,

$$\mathbf{S}_{m,k} := \frac{\mathbf{Z}_{m,k}\mathbf{Z}_{m,k}^H}{\mathrm{Tr}\left(\mathbf{Z}_{m,k}\mathbf{Z}_{m,k}^H\right)} \tag{6.90}$$

Example 6.8.1 (modeling data in wireless networks) We consider using (6.89) to model the big data generated in the wireless networks. At the Wireless Systems Laboratory of Tennessee Technological University, it was recently found that the experimental data for each radio can be modeled as the complex Ginibre matrix. Now consider N such radio receivers. At each time instant t, we have $\mathbf{x}_i \in \mathbb{C}^{N\times 1}, i = 1, \ldots, T$, so that the data matrix is formed as

$$\tilde{\mathbf{X}} = \left(\mathbf{x}_1, \ldots, \mathbf{x}_T\right) \in \mathbb{C}^{N\times T}$$

Consider the singular value equivalent matrix $\mathbf{X} \in \mathbb{C}^{N\times N}$

$$\tilde{\mathbf{X}}\tilde{\mathbf{X}}^H = \mathbf{X}\mathbf{X}^H$$

It is natural to study the product of L such square random matrices $\mathbf{X}_1 \cdots \mathbf{X}_L = \prod_{i=1}^{L} \mathbf{X}_i$, assuming that L such data matrices are observed in the wireless network. Also it is interesting to study the M-th root

$$\left(\mathbf{X}_1 \cdots \mathbf{X}_L\right)^{1/M} = \left(\prod_{i=1}^{L} \mathbf{X}_i\right)^{1/M}$$

□

For a complex quantum system (a system with many degrees of freedom)—such as atoms, nuclei, fundamental particles—it is almost impossible to imagine an exploitable enough theory to compute accurately, for example, the energy levels of such a system. By analogy, we have antenna sensors, smart meters, PMUs, and stocks.

6.9 Products of Rectangular Gaussian Random Matrices

In Example 6.8.1, we encountered rectangular complex random matrices. Below we consider the product of rectangular Gaussian random matrices

$$\mathbf{P} \equiv \mathbf{A}_1\mathbf{A}_2 \cdots \mathbf{A}_L \tag{6.91}$$

of $L \geq 1$ independent rectangular large random Gaussian matrices $\mathbf{A}_\ell, \ell = 1, 2, \ldots, L$, of dimensions $N_\ell \times N_{\ell+1}$. We are interested in the eigenvalue and singular value density of $\mathbf{P}$ in the limit $N_{\ell+1} \to \infty$ and

$$c_\ell \equiv \frac{N_\ell}{N_{L+1}} = \text{finite, for } \ell = 1, 2, \ldots, L+1 \tag{6.92}$$

In other words, all matrix dimensions grow to infinity at fixed rates and, obviously, $c_{L+1} = 1$ The product $\mathbf{P}$ is a matrix of dimensions $N_1 \times N_{L+1}$ and has eigenvalues only

if it is a square matrix: $N_1 = N_{L+1}$. We assume the matrices $\mathbf{A}_\ell$ in the product (6.91) to be complex Gaussian matrices drawn randomly from the ensemble defined by the probability measure

$$d\mu\left(\mathbf{A}_\ell\right) \propto \exp\left(-\frac{\sqrt{N_\ell N_{\ell+1}}}{\sigma_\ell^2}\,\mathrm{Tr}\left(\mathbf{A}_\ell^H \mathbf{A}_\ell\right)\right) D\mathbf{A}_\ell \tag{6.93}$$

where $D\mathbf{A}_\ell \equiv \prod_{a,b} d\left(\mathrm{Re}\left[\mathbf{A}\right]_{ab}\right) d\left(\mathrm{Im}\left[\mathbf{A}\right]_{ab}\right)$ is a flat measure. A normalization constant, fixed by the condition $\int d\mu\left(\mathbf{A}\right) = 1$, is omitted. This is the simplest generalization of the Girko–Ginibre ensemble to rectangular matrices. The σ_ℓ parameters set the scale for the Gaussian fluctuations in $\mathbf{A}_\ell, \ell = 1, \ldots, L$. The entries of each matrix $\mathbf{A}_\ell$ can be viewed as independent centered Gaussian random variables, the variance of the real and imaginary parts being proportional to σ_ℓ^2 and inversely proportional to the square root of the number $N_\ell N_{\ell+1}$ of elements in the matrix.

Let us introduce some notation first. The eigenvalue density $\rho_{\mathbf{X}}\left(\lambda\right)$ of a Hermitian matrix $\mathbf{X}$ is a real function of real argument, while in the case of a non-Hermitian matrix it is a real function of complex argument. In the latter case we write $\rho_{\mathbf{X}}\left(\lambda, \overline{\lambda}\right)$ and treat λ and its conjugate $\overline{\lambda}$ as independent variables.

In the Hermitian case, the eigenvalue density can be computed from a Green's function $G_{\mathbf{X}}\left(z\right)$, which contains the same information as the density itself:

$$\rho_{\mathbf{X}}\left(\lambda\right) = -\frac{1}{\pi}\lim_{\epsilon\to 0^+} \mathrm{Im}\, G_{\mathbf{X}}\left(\lambda + i\epsilon\right) \tag{6.94}$$

For a non-Hermitian matrix, the corresponding Green's function $G_{\mathbf{X}}\left(z, \overline{z}\right)$ is nonholomorphic and therefore we shall write it explicitly as a function of z and $\overline{z}$. In this case the eigenvalue distribution is reconstructed from the Green's function as

$$\rho_{\mathbf{X}}\left(\lambda, \overline{\lambda}\right) = -\frac{1}{\pi}\frac{\partial}{\partial \overline{z}} G_{\mathbf{X}}\left(z, \overline{z}\right) \tag{6.95}$$

It is convenient to use the moment generating function or M transform, which is closely related to the Green's function: $M_{\mathbf{X}}\left(z\right) = zG_{\mathbf{X}}\left(z\right) - 1$. For a Hermitian matrix $\mathbf{X}$ one has

$$M_{\mathbf{X}}\left(z\right) = \sum_{n\geqslant 1} \frac{m_n}{z^n} = \sum_{n\geqslant 1} \frac{1}{z^n} \int \lambda^n \rho_{\mathbf{X}}\left(\lambda\right) d\lambda \tag{6.96}$$

where the m_ns are the moments of the eigenvalue density. If the matrix $\mathbf{X}$ is of finite dimensions $N \times N$, the moments are given by $m_n = \frac{1}{N}\left\langle \mathrm{Tr}\left(\mathbf{X}^n\right)\right\rangle$. The moment-generating function encodes the same information as the Green's function $G_{\mathbf{X}}\left(z\right) = z^{-1} M_{\mathbf{X}}\left(z\right) + z^{-1}$. Thus, one can calculate the corresponding eigenvalue distribution from $M_{\mathbf{X}}\left(z\right)$.

One can also introduce a similar function for non-Hermitian matrices: $M_{\mathbf{X}}\left(z, \overline{z}\right) = zG_{\mathbf{X}}\left(z, \overline{z}\right) - 1$. In this case, however, $M_{\mathbf{X}}\left(z, \overline{z}\right)$ does not play the role of a moment-generating function any more since now one also has mixed moments $\left\langle \mathrm{Tr}\left(\mathbf{X}^n \left(\mathbf{X}^H\right)^k\right)\right\rangle$, which in general depend on the ordering of $\mathbf{X}$ and $\mathbf{X}^H$ in the product under the trace.

The situation is slightly simplified when the M transform is a spherically symmetric function: $M_{\mathbf{X}}\left(z,\overline{z}\right)=\mathcal{M}_{\mathbf{X}}\left(|z|^2\right)$. In this case (6.95) can be cast into the form

$$\rho_{\mathbf{X}}\left(z,\overline{z}\right)=\frac{1}{\pi}\mathcal{M}_{\mathbf{X}}^{'}\left(|z|^2\right)+c\delta^2\left(z,\overline{z}\right) \tag{6.97}$$

where $\mathcal{M}_{\mathbf{X}}^{'}$ is the first derivative of $\mathcal{M}_{\mathbf{X}}$ and $c=1+\mathcal{M}_{\mathbf{X}}(0)$ is a constant representing the fraction of zero modes. In this case the eigenvalue distribution is spherically symmetric as well.

The main result in this section is that the eigenvalue distribution and the M transform of the product (6.91) are *spherically symmetric*. The M transform is shown to satisfy the L-th order polynomial equation

$$\prod_{\ell=1}^{L}\left(\frac{1}{c_\ell}\mathcal{M}_{\mathbf{P}}\left(|z|^2\right)+1\right)=\frac{|z|^2}{\sigma^2} \tag{6.98}$$

where the scale parameter is $\sigma=\sigma_1\sigma_2\cdots\sigma_L$.

An analogous equation for

$$\mathbf{Q}\equiv\mathbf{P}^H\mathbf{P}$$

reads

$$\sqrt{c_1}\frac{M_{\mathbf{Q}}(z)+1}{M_{\mathbf{Q}}(z)}\prod_{\ell=1}^{L}\left(\frac{1}{c_\ell}M_{\mathbf{Q}}(z)+1\right)=\frac{z}{\sigma^2} \tag{6.99}$$

The free argument in (6.98) is $|z|^2$ and z in (6.99).

When $\mathbf{P}$ is a square matrix, then $c_1=1$. When the product of square matrices is considered, all of the $c_\ell,\ell=1,\ldots,L$ become equal to unity and the two equations take the following form:

$$\left(\mathcal{M}_{\mathbf{P}}\left(|z|^2\right)+1\right)^L=\frac{|z|^2}{\sigma^2},\quad M_{\mathbf{Q}}^{-1}(z)\left(M_{\mathbf{Q}}(z)+1\right)^L=\frac{z}{\sigma^2} \tag{6.100}$$

(6.98) can be easily rewritten in terms of the corresponding Green's functions. If one does that and then applies the prescriptions in (6.95) and (6.94), respectively, it becomes clear that

$$\rho_{\mathbf{P}}\left(\lambda,\overline{\lambda}\right)\sim 1/|\lambda|^{2(L-1)/L}\quad\text{and}\ \rho_{\mathbf{Q}}(\lambda)\sim 1/\lambda^{L/(L-1)},\quad\text{as}\ \lambda\to 0 \tag{6.101}$$

In the more general case of rectangular matrices, when solving (6.98) and (6.99) for the Green's functions, one can then see that only those brackets in which $c_1=1$ contribute to the singularity at zero, while all others approach a constant for $z\to 0$. Thus, the eigenvalue density displays the following singularity

$$\rho_{\mathbf{P}}\left(\lambda,\overline{\lambda}\right)\sim 1/|\lambda|^{2(s-1)/s},\quad\text{as}\ \lambda\to 0 \tag{6.102}$$

where s is the number of those ratios among $c_1,\ldots,c_L$, which are exactly equal to unity

$$s\equiv\#\left\{\ell=1,2,\ldots,L:N_\ell=N_{L+1}\right\}=1,2,\ldots,L$$

On the other hand, the eigenvalue density of $\mathbf{Q}$ behaves as

$$\rho_{\mathbf{Q}}(\lambda)\sim 1/\lambda^{-s/(s-1)},\quad\text{as}\ \lambda\to 0 \tag{6.103}$$

where the complementary error function is defined as $\operatorname{erfc}(x) \equiv \left(2/\sqrt{\pi}\right) \int_x^\infty \exp\left(-t^2\right) dt$, and q is a free parameter, whose value is to be adjusted by fitting. (6.103) can be verified numerically.

The third result in this section is a heuristic form for the finite size corrections to the eigenvalue distribution. For a large but finite order of magnitude N of the matrices involved, the eigenvalue distribution is still spherically symmetric. So let $f_N(r)$ denote the radial profile of this distribution, where $r = |\lambda|$. As we shall show, the evolution of the radial shape with the size N is very well described by a simple multiplicative correction:

$$\rho_N(r) \equiv \rho(r) \frac{1}{2} \operatorname{erfc}\left(q(r-\sigma)\sqrt{N}\right) \tag{6.104}$$

In the limit $N \to \infty$ the correction becomes a step function, so that $\rho_\infty(r) = \rho(r)$ for $r \leq \sigma$ and $\rho_\infty(r) = 0$ for $r > \sigma$.

6.10 Product of Complex Wishart Matrices

Example 6.10.1 (product of complex Gaussian matrices) We define the product

$$\mathbf{X}_{r,s} = \mathbf{G}_r \mathbf{G}_{r-1} \cdots \mathbf{G}_1 \left(\tilde{\mathbf{G}}_s \tilde{\mathbf{G}}_{s-1} \cdots \tilde{\mathbf{G}}_1\right)^{-1} \tag{6.105}$$

where the matrices $\tilde{\mathbf{G}}_1, \ldots, \tilde{\mathbf{G}}_s$ have size $N \times N$ and each $\mathbf{G}_k$ is a rectangular standard complex Gaussian matrix of dimension $l_k \times l_{k-1}, l_k \geqslant l_{k-1}$, and $l_0 = N$. For $s = 0$, we have

$$\mathbf{X}_{r,0} = \mathbf{X}_r = \mathbf{G}_r \mathbf{G}_{r-1} \cdots \mathbf{G}_1$$

Our goal is to derive the Stieltjes transform $G(z)$ of $\mathbf{X}_{m,n}^H \mathbf{X}_{m,n}$. We will show that

$$\left(1 - \frac{G(-1/z)}{z}\right)^{r+1} = z\left(\frac{G(-1/z)}{z}\right)^{s+1} \tag{6.106}$$

The eigenvalues of the product Wishart matrix $\mathbf{X}_{r,s}^H \mathbf{X}_{r,s}$ are identical to the eigenvalues of $\left(\mathbf{X}_{r,s-1}^H \mathbf{X}_{r,s-1}\right)\left(\tilde{\mathbf{G}}_s \tilde{\mathbf{G}}_{s-1}\right)^{-1}$. Applying (5.147) and the second equation in (5.164) to the latter, it follows that

$$S_{\mathbf{X}_{r,s}^H \mathbf{X}_{r,s}}(z) = (-z) S_{\mathbf{X}_{r,s-1}^H \mathbf{X}_{r,s-1}}(z)$$

Now iterating this shows

$$S_{\mathbf{X}_{r,s}^H \mathbf{X}_{r,s}}(z) = (-z)^s S_{\mathbf{X}_{r,0}^H \mathbf{X}_{r,0}}(z)$$

For notational convenience, let us now relabel $\mathbf{G}_r \mathbf{G}_{r-1} \cdots \mathbf{G}_1$ to read $\mathbf{G}_1 \mathbf{G}_2 \cdots \mathbf{G}_r$. With this done, we note that $\mathbf{X}_{r,0}^H \mathbf{X}_{r,0}$ has the same eigenvalues as $\left(\mathbf{X}_{r-1,0}^H \mathbf{X}_{r-1,0}\right)\left(\mathbf{G}_r \mathbf{G}_r^H\right)$. Noting that $\mathbf{G}_r \mathbf{G}_r^H$ have the same nonzero eigenvalues as $\mathbf{G}_r^H \mathbf{G}_r$ and applying the first equation in (5.164), (5.147), and iterating we conclude

$$S_{\mathbf{X}_{r,s}^H \mathbf{X}_{r,s}}(z) = \frac{(-z)^s}{(1+z)^r}$$

Recalling (5.163), it follows from this that

$$z = (-1)^s \frac{\left(\Upsilon_{\mathbf{X}_{r,s}^H \mathbf{X}_{r,s}}(z)\right)^{s+1}}{\left(1 + \Upsilon_{\mathbf{X}_{r,s}^H \mathbf{X}_{r,s}}(z)\right)^{r+1}}$$

Recalling (5.161) and performing minor manipulation, (6.106) follows. □

Now let us take a look at the eigenvalue statistics of the product of complex Gaussian matrices for the case $r = s$.

Example 6.10.2 (Eigenvalue statistics of the product of complex Gaussian matrices) We will show that the global density is supported on $(0, \infty)$ and has the explicit form

$$x\rho_{\mathbf{X}_{r,r}^H \mathbf{X}_{r,r}}(x) = \frac{1}{\pi} \frac{x^{1/(r+1)} \sin \frac{\pi}{r+1}}{1 + 2x^{1/(r+1)} \cos \frac{\pi}{r+1} + x^{2/(r+1)}} \tag{6.107}$$

We have from (6.106) in the case $r = s$ that

$$\frac{zG_{\mathbf{X}_{r,r}^H \mathbf{X}_{r,r}}(-z)}{1 - zG_{\mathbf{X}_{r,r}^H \mathbf{X}_{r,r}}(-z)} = z^{1/(r+1)}$$

and thus

$$zG_{\mathbf{X}_{r,r}^H \mathbf{X}_{r,r}}(-z) = 1 - \frac{1}{1 + z^{1/(r+1)}}$$

From the definition (5.157), it follows from this that

$$\int_I \frac{\lambda}{\lambda + z} \rho_{\mathbf{X}_{r,r}^H \mathbf{X}_{r,r}}(\lambda)\, d\lambda = \frac{1}{1 + z^{1/(r+1)}} \tag{6.108}$$

Applying the inverse formula

$$x\rho_{\mathbf{X}_{r,r}^H \mathbf{X}_{r,r}}(x) = -\frac{1}{2\pi i} \left(\frac{1}{1 + z^{1/(r+1)}}\bigg|_{z=xe^{\pi i}} - \frac{1}{1 + z^{1/(r+1)}}\bigg|_{z=xe^{-\pi i}} \right)$$

gives (6.107). □

Example 6.10.3 (singularity of the product of complex Gaussian matrices) Very recently [310], upon the introduction of the variable ϕ according to

$$x = \frac{(\sin(r+1)\phi)^{r+1}}{\sin\phi(\sin r\phi)^r}, \quad 0 < \phi < \frac{\pi}{r+1} \tag{6.109}$$

it has been shown that the corresponding eigenvalue density is given by the succinct expression

$$\rho_{\mathbf{X}_r^H \mathbf{X}_r}(\phi) = \frac{(\sin\phi)^2 (\sin r\phi)^{r-1}}{\pi(\sin(r+1)\phi)^r} \tag{6.110}$$

Of particular interest is the singular behaviour in the original variable x as $x \to 0^+$

$$\rho_{\mathbf{X}_r^H \mathbf{X}_r}(x) \sim \frac{\sin \pi/(r+1)}{\pi x^{r/(r+1)}} \tag{6.111}$$

which follows from (6.109) and (6.110).

Changing variables $\lambda = 1/(1+x)$ transforms the density to have support on (0, 1). It follows from (6.108) that the transformed density satisfies

$$\frac{1}{z}\left(1 - \frac{1}{1+z^{1/(r+1)}}\right) = \int_0^1 \frac{\lambda}{1-(1-z)\lambda} \rho_{\mathbf{X}_r^H \mathbf{X}_r}(\lambda)\, d\lambda$$

and thus that p-th moment is equal to the coefficient of $(1-z)^{p-1}$ in the power series expansion about $z = 1$ of the LHS. The transformed density for $r = s = 1$ is equal to the particular beta density

$$\rho_{\mathbf{X}_1^H \mathbf{X}_1}(\lambda) = \frac{1}{\pi} \frac{1}{\sqrt{\lambda(1-\lambda)}}, \quad 0 < \lambda < 1$$

We remark that the $x \to 0^+$ leading form of (6.107) is exactly the same as that exhibited by (6.111) in the case $s = 0$, suggesting this to be a universal feature valid for general r, s which is independent of s. □

6.11 Spectral Relations between Products and Powers

It is natural to extend (6.91) which is repeated here

$$\mathbf{P} \equiv \mathbf{A}_1 \mathbf{A}_2 \cdots \mathbf{A}_L \tag{6.112}$$

to define the M-th root of some matrix $\mathbf{P}$

$$\mathbf{P}^{1/M} \equiv \mathbf{A}_1^{1/M} \mathbf{A}_2^{1/M} \cdots \mathbf{A}_L^{1/M} \tag{6.113}$$

for an arbitrary non-negative integer $M \geq 1$. $M = 1$ corresponds to the case of (6.91) or (6.112). From Table 6.2, we see that there is a singularity $x^{-k/(k+1)}, k = 0, 1, 2, \ldots$, at the points close to the origin $x \to 0$. After taking the M-th root, we can remove the singularity from the complex plane.

Theorem 6.11.1 ([311, 312]) Consider L identically distributed isotropic matrices $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_L$ generated independently from a given isotropic unitary ensemble (IUE). In the limit $N \to \infty$, the eigenvalue density of the product $\mathbf{P} = \mathbf{X}_1 \mathbf{X}_2 \cdots \mathbf{X}_L$ becomes identical to the eigenvalue density of the L-th power $\mathbf{X}^L$ of a single matrix $\mathbf{X}$ from this ensemble (e.g., $\mathbf{X} = \mathbf{X}_1$).

In other words, the probability that a randomly chosen eigenvalue of $\mathbf{P}$ lies within a circle of radius r: for $N \to \infty$, $\mathbb{P}\left(\lambda_{\mathbf{P}} < r\right)$ approaches $\mathbb{P}\left(\lambda_{\mathbf{X}}^L < r\right)$, which is the probability that a randomly chosen eigenvalue of $\mathbf{X}^L$ lies within the same circle.

One can use this above observation to derive the eigenvalue density of the product $\mathbf{P} = \mathbf{X}_1 \mathbf{X}_2 \cdots \mathbf{X}_L$ if the eigenvalue density of $\mathbf{X}$ is known. In particular one can immediately

Table 6.2 A normalized Wishart-like matrix is defined as $\mathbf{S} = \mathbf{Z}\mathbf{Z}^H / \mathrm{Tr}\left(\mathbf{Z}\mathbf{Z}^H\right)$. Random matrix $\mathbf{Z}$ defined in (6.89) is constructed out of random unitary matrices $\mathbf{U}_i, i = 1, \ldots, m$ distributed according to the Haar measure and/or (independent) random Ginibre matrices $\mathbf{G}_j, j = 1, \ldots, k$ of a given size N. Asymptotic distribution $P(x)$ of the density of a rescaled eigenvalue $x = N\lambda$ of ρ for $N \to \infty$ is characterized by the singularity at 0, its support $[a, b]$, the second moment M_2 determining the average purity $\left\langle \mathrm{Tr}\left(\mathbf{S}^2\right)\right\rangle = M_2/N$ and the mean entropy $\int_a^b -x \ln x P(x) dx$, according to which the table is ordered. Taken from [309].

m	k	Matrix W	Distribution $P(x)$	Singularity at $x \to 0$	Support $[a, b]$	M_2	Mean entropy
1	0	$\mathbf{U}_1$	$\delta(1) = \pi^{(0)}(x)$	—	$\{1\}$	1	0
2	0	$\mathbf{U}_1 + \mathbf{U}_2$	arcsin	$x^{-1/2}$	$[0, 2]$	3/2	$\ln 2 - 1 \approx -0.307$
3	0	$\mathbf{U}_1 + \mathbf{U}_2 + \mathbf{U}_3$	—	$x^{-1/2}$	$[0, 2\frac{2}{3}]$	5/3	≈ -0.378
4	0	$\mathbf{U}_1 + \mathbf{U}_2 + \mathbf{U}_3 + \mathbf{U}_4$	—	$x^{-1/2}$	$[0, 3]$	7/8	≈ -0.4111
1	1	$\mathbf{G} \sim \mathbf{U}G$	Marchenko–Pastur $\pi^{(1)}$	$x^{-1/2}$	$[0, 4]$	2	$-1/2 = 0.5$
2	1	$(\mathbf{U}_1 + \mathbf{U}_2)\mathbf{G}$	Bures	$x^{-2/3}$	$[0, 3\sqrt{3}]$	5/2	$-\ln 2 \approx -0.693$
1	2	$\mathbf{G}_1\mathbf{G}_2$	Fuss–Catalan $\pi^{(2)}$	$x^{-2/3}$	$[0, 6\frac{3}{4}]$	3	$-5/6 \approx -0.833$
1	..	...	...	...	...	...	...
1	k	$\mathbf{G}_1 \cdots \mathbf{G}_k$	Fuss–Catalan $\pi^{(k)}$	$x^{-k/(k+1)}$	$[0, (k+1)^{k+1}/k^k]$	$k+1$	$-\sum_{j=2}^{k+1} \frac{1}{j}$

show that the eigenvalue distribution of the product of L independent Girko–Ginibre matrices has a simple form:

$$\rho\left(z,\overline{z}\right)=\frac{1}{\pi L}|z|^{-2+2/L}\quad\text{for}\ \ |z|\leqslant 1 \tag{6.114}$$

and zero for $|z|>1$. The matrix $\mathbf{PP}^H$ obtained from the product $\mathbf{P}$ of the L Girko–Ginibre matrices generate a Fuss–Catalan family of distributions that have, however, a much more complicated limiting eigenvalue density.

Theorem 6.11.1 is a counterintuitive result, so let us stress that it only holds in the limit $N\to\infty$. Now let us derive the results in Theorem 6.11.1. We emphasize the approach used here.

For an R-diagonal (isotropic) matrix $\mathbf{X}$ given by the radial decomposition $\mathbf{X}=\mathbf{HU}$, where $\mathbf{H}$ Hermitian and $\mathbf{U}$ is a Haar unitary matrix, the two matrices $\mathbf{XX}^H=\mathbf{H}^2$ and $\mathbf{X}^H\mathbf{X}=\mathbf{U}^H\mathbf{H}^2\mathbf{U}$ have identical eigenvalues and therefore the S-transforms for $\mathbf{X}^H\mathbf{X}$ and $\mathbf{XX}^H$ are identical:

$$S_{\mathbf{XX}^H}(z)=S_{\mathbf{X}^H\mathbf{X}}(z)=S_{\mathbf{H}^2}(z) \tag{6.115}$$

Consider an isotropic unitary ensemble of random matrices $\mathbf{X}=\mathbf{HU}\in\mathbb{C}^{N\times N}$. In the large N limit the random matrices can be represented as free random variables and one can use the Haagerup–Larsen theorem [292] that relates the eigenvalue density of $\mathbf{X}$ to the eigenvalue density of $\mathbf{H}^2$ by the following formula:

$$S_{\mathbf{H}^2}\left(F_{\mathbf{X}}(r)-1\right)=\frac{1}{r^2} \tag{6.116}$$

where $F_{\mathbf{X}}(r)$ is the cumulative density function for the density of eigenvalues of $\mathbf{X}$ on the complex plane and $S_{\mathbf{H}^2}(z)$ the S-transform for the matrix $\mathbf{H}^2$. The cumulative density function

$$F_{\mathbf{X}}(r)=\int_{|z|\leqslant r}\rho_x\left(z,\overline{z}\right)d^2z=2\pi\int_0^r s\rho_x(s)\,ds=\int_0^r p_x(s)\,ds \tag{6.117}$$

can be interpreted as the fraction of eigenvalues of $\mathbf{X}$ in the circle of radius r centered at the origin of the complex plane. It is related to the eigenvalue density $\rho_x\left(z,\overline{z}\right)=\rho_x(|z|)$ that depends on the distance from the origin $|z|$. The integrand $p_x(s)\,ds=2\pi s\rho_x(s)\,ds$ is interpreted as the probability of finding eigenvalues of $\mathbf{X}$ in a narrow ring of radii $|z|$ and $|z|+d|z|$:

$$F_{\mathbf{X}}'(r)=p_x(r)=2\pi r\rho_x(r) \tag{6.118}$$

The prime denotes the derivation with respect to the radial variable. The cumulative density function $F_{\mathbf{X}}(r)$ enters equation (6.116) as an argument of the S-transform $S_{\mathbf{H}^2}(z)$ that is related to the eigenvalue density $\rho_{\mathbf{H}^2}(\lambda)$ of the matrix $\mathbf{X}^2$. The Haagerup–Larsen theorem states also that the support of the eigenvalue density of $\mathbf{X}$ is a ring of radii $R_{\min}$ and $R_{\max}$ or a disk (if $R_{\min}=0$)

$$R_{\min}^2=\int_0^\infty\lambda^{-1}\rho_{\mathbf{H}^2}(\lambda)\,d\lambda,\quad R_{\max}^2=\int_0^\infty\lambda\rho_{\mathbf{H}^2}(\lambda)\,d\lambda \tag{6.119}$$

It follows from (6.115) that (6.116) can be rewritten as

$$S_{\mathbf{X}^H\mathbf{X}}\left(F_{\mathbf{X}}(r)-1\right)=\frac{1}{r^2} \tag{6.120}$$

Now we are in a position to apply (6.120) to the product of L identically distributed R-diagonal (isotropic) matrices $\mathbf{P}_L = \mathbf{X}_1\mathbf{X}_2\cdots\mathbf{X}_L$. The resulting matrix has an identical eigenvalues as $\mathbf{H}_L\mathbf{U}_L$, where $\mathbf{H}_L^2 = \mathbf{P}_L^H\mathbf{P}_L$ so we can apply (6.120) replacing in this equation $\mathbf{X}$ by $\mathbf{P}_L$:

$$S_{\mathbf{P}_L^H\mathbf{P}_L}\left(F_{\mathbf{P}_L}(r)-1\right) = \frac{1}{r^2} \tag{6.121}$$

The S-transform for the matrix $\mathbf{P}_L^H\mathbf{P}_L$ that appears in (6.121) can be substituted by the S-transforms for individual terms in the product. Indeed, writing

$$\mathbf{P}_L^H\mathbf{P}_L = \mathbf{X}_L^H\mathbf{P}_{L-1}^H\mathbf{P}_{L-1}\mathbf{X}_L \tag{6.122}$$

where $\mathbf{P}_{L-1} = \mathbf{X}_1 \ldots \mathbf{X}_{L-1}$ we find that

$$S_{\mathbf{P}_L^H\mathbf{P}_L} = S_{\mathbf{P}_{L-1}^H\mathbf{P}_{L-1}}S_{\mathbf{X}_L^H\mathbf{X}_L} \tag{6.123}$$

because, due to the cyclic properties of trace, the moments of $\mathbf{X}_L^H\mathbf{P}_{L-1}^H\mathbf{P}_{L-1}\mathbf{X}_L$ are identical to those of $\mathbf{X}_L\mathbf{X}_L^H\mathbf{P}_{L-1}^H\mathbf{P}_{L-1}$ and the moments of $\mathbf{X}_L\mathbf{X}_L^H$ are identical to those of $\mathbf{X}_L^H\mathbf{X}_L$. Applying (6.123) recursively we eventually obtain

$$S_{\mathbf{P}_L^H\mathbf{P}_L} = \prod_{i=1}^{L} S_{\mathbf{X}_i^H\mathbf{X}_i} \tag{6.124}$$

Taking into account that all $\mathbf{X}_i$ are identically distributed and having the same S-transform (that we denote by $S_{\mathbf{X}^H\mathbf{X}}$) we can write (6.124) as

$$S_{\mathbf{P}_L^H\mathbf{P}_L} = S_{\mathbf{X}^H\mathbf{X}}^L \tag{6.125}$$

Inserting this into (6.121) we have

$$S_{\mathbf{X}^H\mathbf{X}}\left(F_{\mathbf{P}_L}(r)-1\right) = \frac{1}{r^{2/L}} \tag{6.126}$$

(6.126) has a form identical to (6.120) except that on the left hand side $F_{\mathbf{X}}(r)$ is replaced by $F_{\mathbf{P}_L}(r)$ and on the right hand side r is replaced by $r^{1/L}$. From this observation it immediately follows that

$$F_{\mathbf{P}_L}(r) = F_{\mathbf{X}}\left(r^{1/L}\right) = F_{\mathbf{X}^L}(r) \tag{6.127}$$

(6.127) follows from the fact that the eigenvalues of $\mathbf{X}^L$ are equal to the L-th power of the corresponding eigenvalues of $\mathbf{X}$:

$$F_{\mathbf{X}^L}(r) \equiv \mathbb{P}\left(|\lambda|^L \leqslant r\right) = \mathbb{P}\left(|\lambda| \leqslant r^{1/L}\right) \equiv F_{\mathbf{X}}\left(r^{1/L}\right) \tag{6.128}$$

So we see that the product of L identically distributed isotropic matrices $\mathbf{P}_L = \mathbf{X}_1\mathbf{X}_2\cdots\mathbf{X}_L$ has the same eigenvalue distribution as the L-th power $\mathbf{X}^L$ of a single matrix $\mathbf{X}$ in the product. In practice, the eigenvalue distribution of $\mathbf{P}_L$ can be calculated directly from the eigenvalue distribution of a single matrix $\mathbf{X}$ by substituting $r \to r^{1/L}$ in the cumulative distribution function $F_{\mathbf{X}}(r)$ (6.117). The corresponding eigenvalue densities may be found using (6.117). They read

$$p_{\mathbf{P}_L}(r) = \frac{1}{L}r^{1/L-1}p_{\mathbf{X}}\left(r^{1/L}\right) \tag{6.129}$$

and

$$\rho_{\mathbf{P}_L}(r) = \frac{1}{L}r^{2/L-2}\rho_{\mathbf{X}}\left(r^{1/L}\right) \tag{6.130}$$

Example 6.11.2 (Girko–Ginibre matrices) Girko-Ginibre matrices $\mathbf{X}$ have a uniform distribution $\rho_{\mathbf{X}}(r) = 1/\pi$ inside the unit circle $|z| \leq 1$. We have

$$F_{\mathbf{X}}(r) = 2\int_0^r ydy = r^2 \quad \text{for } r \leqslant 1 \tag{6.131}$$

and 1 otherwise. For the product of L-independent Girko–Ginibre matrices we have (6.127)

$$F_{\mathbf{P}_L}(r) = r^{2/L} \quad \text{for} \quad r \leq 1 \tag{6.132}$$

and one otherwise. Taking the derivative with respect to r (6.118), we find the corresponding densities:

$$p_{\mathbf{P}_L}(r) = \frac{2}{L} r^{2/L-1}\theta(1-r)$$

and

$$\rho_{\mathbf{P}_L}(r) = \frac{2}{\pi L} r^{2/L-2}\theta(1-r)$$

where θ denotes the Heaviside step function. □

6.12 Products of Finite-Size I.I.D. Gaussian Random Matrices

Products of matrices lose much of the symmetry of the individual matrices and are generically complex. For example a product of symmetric matrices will not be symmetric in general. For simplicity, we will look at matrices with a minimum of symmetry.

A striking property of RMT is its universality, which is the independence of the underlying distribution of the individual matrix elements. It is usually manifest in the limit of large matrix size. However, if we study the local, microscopic behavior of the spectrum on the scale of the mean level spacing between singular values, it is often vital to have a detailed knowledge of the joint distribution of singular values (or eigenvalues) at hand for finite matrix size.

We consider the product of L complex non-Hermitian, independent random matrices, each of size $N \times N$ with independent identically distributed Gaussian entries (Ginibre matrices). We compute all eigenvalue density correlation functions exactly for finite N and fixed L. Given the product $\mathbf{P}_L$ of L independent matrices $\mathbf{X}_i, i = 1, \ldots, L$, each of size $N \times N$ drawn from the Ginibre ensemble with Gaussian distribution proportional to $\exp\left[-\operatorname{Tr}\mathbf{X}_i^H\mathbf{X}_i\right]$

$$\mathbf{P}_L = \mathbf{X}_1\mathbf{X}_2\cdots\mathbf{X}_L. \tag{6.133}$$

The $L = 1$ case is the Ginibre ensemble, while the $L = 2$ case is the Wishart ensemble.

The partition function Z_L can be expressed as an integral of the joint probability distribution function $\mathcal{P}_{jpdf}$ of the complex eigenvalues $z_i, i = 1, \ldots, N$, are given by

$$Z_L = C_L \int \prod_{a=1}^{N} d^2 z_a w_L(z_a) \prod_{b>a}^{N} |z_b - z_a|^2 \equiv \int \prod_{a=1}^{N} d^2 z_a \mathcal{P}_{jpdf}(\{z\}) \tag{6.134}$$

where C_L is some known constant. The weight function $w_L(z)$ that depends only on the modulus $|z|$ is given by the so-called Meijer G-function. The corresponding kernel of polynomials orthonormal with respect to that weight reads

$$K_N^{(L)}\left(z_i, z_j\right) = \sqrt{w_L\left(z_i\right) w_L\left(z_j\right)} \sum_{k=0}^{N-1} \frac{1}{(\pi k!)^L} \left(z_i z_j^*\right)^k \tag{6.135}$$

The k-point density correlation functions then easily follow to be the determinant of that kernel

$$\begin{aligned} R_k^{(L)}\left(z_1, \ldots, z_k\right) &\equiv \tfrac{N!}{(N-k)!} \tfrac{1}{Z_L} \int \mathcal{P}_{ipdf}\left(\{z\}\right) d^2 z_{k+1} \cdots d^2 z_N \\ &= \det_{1 \leqslant i,j \leqslant k} \left[K_N^{(L)}\left(z_i, z_j\right)\right] \end{aligned} \tag{6.136}$$

For large N and large arguments $|z| >> 1$ the eigenvalue density behaves as

$$R_1^{(L)}\left(z_1, \ldots, z_k\right) = K_N^{(L)}(z, z) \approx \frac{|z|^{\frac{2}{L}-2}}{L\pi} \frac{1}{2} erfc\left(\frac{\sqrt{L}\left(|z|^{2/L} - N\right)}{\sqrt{2}|z|^{1/L}}\right) \tag{6.137}$$

By zooming into the region around the edge of the support, which is $z \approx N^{L/2}$ in (6.137), we obtain

$$R_1^{(L)}\left(z_1, \ldots, z_k\right) = K_N^{(L)}(z, z) \approx \frac{|z|^{\frac{2}{L}-2}}{L\pi} \frac{1}{2} \operatorname{erfc}\left(\frac{\sqrt{L}\left(|z|^{2/L} - N\right)}{\sqrt{2}|z|^{1/L}}\right) \tag{6.138}$$

This result depending only on the radial distance from the edge is universal in the sense that it is valid for all L. It is convenient to recast this result (6.138) into a rescaled density with compact support that is normalized to unity. Using the rescaled variable $w = zN^{-L/2}$, we define the following density, for which the radius of the eigenvalue support approaches 1 for $N \to \infty$:

$$\begin{aligned} \rho_L(w) &\equiv \lim_{N>>1} \tfrac{1}{N} N^L R_1^{(L)}\left(N^{L/2} w\right) = \tfrac{|w|^{\frac{2}{L}-2}}{L\pi} \tfrac{1}{2} \operatorname{erfc}\left(\tfrac{\sqrt{LN}\left(|z|^{2/L}-1\right)}{\sqrt{2}|z|^{1/L}}\right) \\ &= \tfrac{|w|^{\frac{2}{L}-2}}{L\pi} \tfrac{1}{2} \operatorname{erfc}\left(\sqrt{\tfrac{2N}{L}}\left(|w|-1\right)\right) \end{aligned} \tag{6.139}$$

The complementary error function changes only in a narrow strip around the unit circle $|w| = 1$, of a width proportional to $1/\sqrt{N}$. We see that the width of the crossover region around the edge$|w| = 1$ is proportional to $\sqrt{L}$, the square root of the number of multiplied matrices.

It is instructive to compare this result (6.139) with the limiting density for large-N of the L-th power $\mathbf{X}^L$ of a single Ginibre matrix $\mathbf{X}$. This density is given by exactly the same distribution, however with a different dependence on L:

$$\rho_L(w) = \frac{|w|^{\frac{2}{L}-2}}{L\pi} \frac{1}{2} \operatorname{erfc}\left(\frac{\sqrt{2N}}{L}\left(|w|-1\right)\right) \tag{6.140}$$

Here, the width of the crossover region is proportional to L and not to $\sqrt{L}$. In a sense, the finite size corrections are stronger for the L-th power than for the product of L independent Ginibre matrices.

From (6.139), the mean or macroscopic large-N density can be obtained:

$$\rho_{macro}^{(L)}(w) \equiv \lim_{N\to\infty} \frac{1}{N} N^L R_1^{(L)}\left(z = N^{L/2} w\right) = \frac{|w|^{\frac{2}{L}-2}}{L\pi} \Theta(1-|w|) \tag{6.141}$$

where Θ is Heaviside's function.

The above treatment is only valid for the products of finite-size square matrices. We can extend this discussion to include products of rectangular matrices. In particular, we consider the product matrix

$$\mathbf{Y}_L = \mathbf{X}_L \mathbf{X}_{L-1} \cdots \mathbf{X}_1 \tag{6.142}$$

where $\mathbf{X}_\ell$ are $N_\ell \times N_{\ell-1}$ real $\beta = 1$, complex $\beta = 2$, and quaternion ($\beta = 4$) matrices from the Wishart ensemble. We deal with the singular values of such matrices, and the spectral correlation functions of $\mathbf{Y}_L \mathbf{Y}_L^H$.

6.13 Lyapunov Exponents for Products of Complex Gaussian Random Matrices

The application that we are considering here is the time-varying topology of a large network (e.g. wireless communications, power grids). Our aim is to combine the contemporary interest in the eigenvalues of large random matrices with the topic of products of random matrices by studying eigenvalue distributions of products of random matrices where the size of the matrices is large. Lyapunov exponents are useful tools that measure the sensitivity of a dynamical system with respect to initial conditions. Let $f : X \to X$ be a differentiable map of a manifold X to itself. The dependence on small perturbations in the initial conditions can be measured by the growth of matrix products $\mathbf{P}_n = \mathbf{A}_n \mathbf{A}_{n-1} \cdots \mathbf{A}_1$, where $\mathbf{A}_k := f'\left(\mathbf{x}_k\right)$ and $\mathbf{x}_k = f\left(\mathbf{x}_{k-1}\right)$. In order to quantify what is meant by a typical initial position, the manifold X is usually endowed with a probability measure. Then, $\mathbf{A}_k$ are random matrices and therefore we are led to the study of random matrix products.

Let

$$\mathbf{P}_n = \mathbf{A}_n \mathbf{A}_{n-1} \cdots \mathbf{A}_1 \tag{6.143}$$

where each $\mathbf{A}_i$ is a $d \times d$ independent, identically distributed random matrix such that the diagonal elements of $\mathbf{A}^H \mathbf{A}$ have finite second moments. According to the multiplicative ergodic theorem of Oseledec [313], the limiting matrix

$$\mathbf{V}_d := \lim_{n\to\infty} \left(\mathbf{P}_n^H \mathbf{P}_n\right)^{1/(2n)} \tag{6.144}$$

is well defined, with d positive real eigenvalues $e^{\mu_1} \geqslant e^{\mu_2} \geqslant \cdots \geqslant e^{\mu_d}$. The $\{\mu_i\}$ are referred to as the Lyapunov exponents.

The key fact about Lyapunov exponents is that they satisfy the following relation:

$$\mu_1 + \cdots + \mu_k = \sup \lim_{n\to\infty} \frac{1}{n} \log \mathrm{Vol}_k\left(\mathbf{y}_1(n), \ldots, \mathbf{y}_k(n)\right) \tag{6.145}$$

where $\mathbf{y}_i(n) = \mathbf{P}_n \mathbf{y}_i(0)$ and the supremum is over all choices of linearly independent vectors $\mathbf{y}_i(0)$. It can be proved that the supremum is in fact not needed in this formula. In words, the sum of k largest Lyapunov exponents measures the average growth rate in the volume of a k-dimensional element when we apply linear transformations

specified by matrices $\mathbf{A}_i$. If these matrices are independent and Gaussian, then this formula can be significantly simplified. Namely, let $\mathbf{G}^{(i)}$ be independent random $d \times d$ matrices whose entries are independent (real) standard Gaussian entries, and $\mathbf{\Sigma}^{1/2}$ is a (real) positive-definite $d \times d$ matrix. Let $\mathbf{A}_i = \mathbf{\Sigma}^{1/2}\mathbf{G}^{(i)}$. We will call these matrices real Gaussian matrices with covariance matrix $\mathbf{\Sigma}$. The crucial observation is that the distribution of $\mathbf{A}_i^T\mathbf{A}_i$ is invariant relative to the transformation

$$\mathbf{A}_i^T\mathbf{A}_i \to \mathbf{Q}^T\mathbf{A}_i^T\mathbf{A}_i\mathbf{Q},$$

where $\mathbf{Q}$ is an arbitrary orthogonal matrix. This implies that the changes in the volume of a k-dimensional element are independent from step to step and that their distribution is the same as if they were applied to the element spanned by the standard basis vectors $\mathbf{e}_i$,

$$\begin{aligned}\mu_1 + \cdots + \mu_k &= \mathbb{E}\log \mathrm{Vol}_k\left(\mathbf{A}_1\mathbf{e}_1, \ldots, \mathbf{A}_1\mathbf{e}_k\right)\\ &= \frac{1}{2}\mathbb{E}\log\det\left(\mathbf{G}_k^T\mathbf{\Sigma}\mathbf{G}_k\right),\end{aligned} \tag{6.146}$$

where $\mathbf{G}_k$ denotes a random $d \times k$ matrix with the identically distributed standard Gaussian entries (see [314] for details). Sometimes it is useful to write this formula as

$$\mu_1 + \cdots + \mu_k = \frac{1}{2}\frac{d}{d\mu}\mathbb{E}\left[\det\left(\mathbf{G}_k^T\mathbf{\Sigma}\mathbf{G}_k\right)^{\mu}\right]\Big|_{\mu=0} \tag{6.147}$$

This argument works for complex Gaussian matrices as well.

While formula (6.144) allows one to compute all Lyapunov exponents, it is essentially a multidimensional integral, which can be computationally demanding. For this reason, it is of interest to obtain a more explicit method for the Lyapunov exponent calculation. For big-data applications, real-time computing is required.

Implicit in the need for efficient computational methods is that it is generally not possible to compute the Lyapunov exponents analytically. Some noteworthy exceptions occur in the case $d = 2$. For general d, except from the case of diagonal matrices, it seems that the only exact computation of the Lyapunov exponents recorded in the literature is when the $\mathbf{A}_i$ are real Gaussian matrices with entries independent standard real normals. For real Gaussian matrices and the simplest situation when $\mathbf{\Sigma} = \sigma^2\mathbf{I}$ and $\mathbf{I}$ is the identity matrix, Newman [314] found that

$$\mu_i = \frac{1}{2}\left(\log\left(2\sigma^2\right) + \Psi\left(\frac{d-i+1}{2}\right)\right) \quad (i = 1, \ldots, d) \tag{6.148}$$

where $\Psi(x)$ denotes the digamma function, $\Psi(x) := \left(\log\Gamma(x)\right)'$. At the positive integer points, $\Psi(n) = \sum_{k=1}^{n-1}\frac{1}{k} - \gamma$, where $\gamma = 0.5772...$ is the Euler constant. At half-integers, $\Psi(n+1/2) = \sum_{k=1}^{n-1}\frac{1}{k-1/2} - 2\log 2 - \gamma$. The asymptotic behavior of the digamma function is given by the formula $\Psi(z) = \log z - \frac{1}{2z} - \frac{1}{12z^2}\left(1 + O\left(\frac{1}{z^2}\right)\right)$. In particular if we normalize $\sigma^2 = 1/d$, then for $d = 1$ the largest Lyapunov exponent $\mu_1 = \left[-\log 2 - \gamma\right]/2$ and $d \to \infty$, $\mu_1 = -\frac{1}{2d} + O\left(\frac{1}{d^2}\right)$. Another explicit formula is known for the sum of all Lyapunov exponents. Indeed, if $k = d$, then $\det\left(\mathbf{G}_k^T\mathbf{\Sigma}\mathbf{G}_k\right) = \det\left(\mathbf{G}_k^T\mathbf{G}_k\right)\det(\mathbf{\Sigma})$, and therefore formula (6.147) becomes

$$\mu_1 + \cdots \mu_d = \frac{1}{2}\log\det\mathbf{\Sigma} + \frac{d}{d\mu}\mathbb{E}\left[\det\left(\mathbf{G}_d\mathbf{G}_d^T\right)^{\mu}\right]\Big|_{\mu=0}$$

In [315], Forrester showed that this implies that

$$\mu_1 + \cdots \mu_d = \frac{1}{2}\sum_{i=1}^{d}\left(\log\left(\frac{2}{y_i}\right) + \Psi\left(\frac{i}{2}\right)\right) \tag{6.149}$$

where y_i are eigenvalues of $\mathbf{\Sigma}^{-1}$.

Forrester has also established analog of formulas (6.148) and (6.149) for the complex-valued Gaussian matrices. Recall that in general the density for a Gaussian matrix $\mathbf{A}$ with covariance matrix $\mathbf{\Sigma}$ is given by

$$\mathbb{P}(\mathbf{A}) = c_\beta \det\left(\mathbf{\Sigma}^{-k}\right) \exp\left[-\frac{\beta}{2}\operatorname{Tr}\left(\mathbf{A}^T\right)\mathbf{\Sigma}^{-1}\mathbf{A}\right]$$

where $\beta = 1, 2$, or 4 for real, complex or quaternion matrices and c_β is a normalization constant. Equivalently, $\mathbf{A}$ can be obtained as $\mathbf{\Sigma}^{1/2}\mathbf{G}$, and $\mathbf{G}$ is a real, complex or quaternion matrix with independent entries. The entries of $\mathbf{G}$ have components that are real Gaussian variables with variance $1/\beta$. Namely, Forrester showed that in the case of the complex-valued Gaussian matrices with $\mathbf{\Sigma} = \sigma^2\mathbf{I}$

$$2\mu_i = \log\sigma^2 + \Psi(d - i + 1) \tag{6.150}$$

(See Proposition 1 in [315] and note that the absence of $1/2$ before Ψ is a typo.)

If $\sigma^2 = 1/d$, then for $d = 1$ the largest Lyapunov exponent $\mu_1 = -\gamma/2$ and for $d \to \infty$, $\mu_1 = -\frac{1}{d} + O\left(\frac{1}{d^2}\right)$. The sum rule in the complex valued case is

$$\mu_1 + \cdots \mu_d = \frac{1}{2}\sum_{i=1}^{d}\left(\log\left(\frac{1}{y_i}\right) + \Psi(i)\right)$$

A significant advance that Forrester achieved in the complex-valued case is an explicit formula for all Lyapunov exponents valid in the case of general Σ. Namely, it is shown in [315] that

$$\mu_k = \frac{1}{2}\Psi(k) + \frac{1}{2\prod_{i<j}(y_i - y_j)} \det\begin{bmatrix} \left[y_j^{i-1}\right]_{i=1,\ldots,k-1;j=1,\ldots,d} \\ \left[(\log y_j)\, y_j^{k-1}\right]_{j=1,\ldots,d} \\ \left[y_j^{i-1}\right]_{i=k+1,\ldots,d;j=1,\ldots,d} \end{bmatrix} \tag{6.151}$$

where y_i are eigenvalues of $\mathbf{\Sigma}^{-1}$. In particular, for $k = 1$, one can rewrite this as

$$\mu_1 = \frac{1}{2}\left[\Psi(1) - \sum_{j=1}^{d}\frac{\log y_i}{\prod_{\ell\neq j}\left(1 - y_j/y_\ell\right)}\right]$$

provided that all y_i are different.

The proof of formula (6.151) is based on the Harish–Chandra–Itzykson–Zuber integral and cannot be directly generalized to the case of real or quaternion Gaussian matrices.

In fact, it appears that for the real-valued case with general Σ, an explicit formula (from [316]) is only known for products of 2×2 Gaussian matrices:

$$\mu_1 = \frac{1}{2}\left[\Psi(1) + \log\left(\frac{1}{2}\operatorname{Tr}\mathbf{\Sigma} + \sqrt{\det\mathbf{\Sigma}}\right)\right] \tag{6.152}$$

Some explicit formulas are also known for 2×2 random matrices with non-Gaussian entries, see [317]. In addition, there are methods that sometime allow one to compute Lyapunov exponents efficiently even when explicit formulas are not available, (see [318]).

The theorem below is to derive an explicit formula for the largest Lyapunov exponent that would be applicable in the real and quaternion-valued case with general $\mathbf{\Sigma}$.

Theorem 6.13.1 ([319]) Let $\mathbf{A}_i$ be independent Gaussian matrices with covariance matrix $\mathbf{\Sigma}$. Let the entries be real, complex or quaternion, according to whether $\beta = 1; 2;$ or 4. Assume that the eigenvalues of $\mathbf{\Sigma}$ are $\sigma_i^2 = 1/y_i$. Then, the following formula holds for the largest Lyapunov exponent of $\mathbf{A}_i$

$$2\mu_1 = \Psi(1) + \log\left(\frac{2}{\beta}\right) + \int_0^\infty \left[\mathbb{I}_{[0,1]}(x) - \prod_{i=1}^{d}\left(1 + \frac{x}{y_i}\right)^{-\beta/2}\right]\frac{1}{x}dx \tag{6.153}$$

Consider a model with a spike. Assume that all $y_i = 1$, for $i = 1, \dots, d-1$ and $y_d = 1/\theta < 1$. This means that the covariance matrix $\mathbf{\Sigma}$ has a spike $\theta > 1$; or informally that one of the rows in matrices $\mathbf{A}_i$ has the size which is $\sqrt{\theta}$ larger than other rows. We ask the question about the behavior of the largest Lyapunov exponent when d; or θ; or both, are large. Assume first that $\beta = 2$. We can write

$$\begin{aligned} 2\mu_1 &= \Psi(1) + \int_0^\infty \left[\mathbb{I}_{[0,1]}(x) - \frac{1}{(1+x)^{d-1}(1+\theta x)}\right]\frac{1}{x}dx \\ &= \Psi(d) + f_d \end{aligned}$$

where

$$f_d = (\theta - 1)\int_0^\infty \frac{1}{(1+x)^{d-1}(1+\theta x)} \leqslant \frac{\theta - 1}{d} \tag{6.154}$$

Hence, if $\theta = O(d)$ and $d \to \infty$, then

$$2\mu_1 \sim \Psi(d) \sim \log d.$$

In other words, in this case the spike θ in $\mathbf{\Sigma}$ cannot influence the leading order asymptotics of the largest Lyapunov exponent.

It is still interesting to find out what is the contribution of the spike θ to the Lyapunov exponent even though it is of a lower order than the leading asymptotics. (Indeed, the leading term asymptotics can be removed if we rescale all the entries in the matrices $\mathbf{A}_i$ by $\sigma = 1/\sqrt{d}$.)

Theorem 6.13.2 ([319]) Suppose that $\mathbf{A}_i$ are independent $d \times d$ Gaussian matrices with the covariance matrix $\mathbf{\Sigma}$ and that the eigenvalues of $\mathbf{\Sigma}$ are $\sigma_i^2 = 1$, for $i = 1, \dots, d-1$ and $\sigma_i^2 = \theta > 1$. Let $\theta = d/t$, where $0 < t < d$. In the complex case ($\beta = 2$), we have the following estimate

$$\begin{aligned} 2\mu_1 &= \log d + e^t \int_1^\infty e^{-tx}\frac{dx}{x} + O_t(1/d) \\ &= \log d - e^t \operatorname{Ei}(-t) + O_t(1/d) \end{aligned} \tag{6.155}$$

where $\mathrm{Ei}(x)$ is the exponential integral function. In the real case ($\beta = 1$)

$$2\mu_1 = \log d + e^{t/2} \int_1^\infty e^{-tx/2} \frac{1}{\sqrt{x}\left(\sqrt{x}+1\right)} dx + O_t(1/d)$$

The above theorem implies the following: for the largest Lyapunov exponent μ_1, we have

$$\lim_{d\to\infty} d\left(\mu_1 - \log d\right) = \theta - \frac{3}{2}$$

When d is fixed and θ goes to infinity, we have

$$2\mu_1 \sim \log\theta - \gamma$$

6.14 Euclidean Random Matrices

A special class of random matrices are the so-called Euclidean random matrices. See also Section 16.1.5 for its connection with random geometric graphs. The elements A_{ij} of an $N \times N$ Euclidean random matrix $\mathbf{A}$ are given by a *deterministic* function f of positions of pairs of points that are randomly distributed in a finite region V of Euclidean space:

$$A_{ij} = f\left(\mathbf{r}_i, \mathbf{r}_j\right), \quad i,j = 1,\ldots,N$$

Here, the N points $\mathbf{r}_i$ are randomly distributed inside some region V of the d- dimensional Euclidean space with a uniform density $\rho = N/V$. In general, the random matrix $\mathbf{A}$ is non-Hermitian.

This model may be applied to massive MIMO where each antenna is viewed as a scattering center located at a random position $\mathbf{r}_i, i = 1,\ldots,N$. We are interested in the collective radiation from the region V containing N randomly located antennas, especially when N is large, say $N = 10^4$. This is analogous with collective spontaneous emission in dense atomic systems consisting of N atoms [320]. This model of three-dimensional region of space is of interest to unmanned aerial vehicles (UAVs). One extension of the work in this section is to consider the impact of multipath on the eigenvalue distributions, as only free-space Green's function is considered for the path with the line of sight (LOS) between the transmitter and the receiver.

For arbitrary V, we have

$$\mathbf{A} = \mathbf{H}\mathbf{T}\mathbf{H}^H \tag{6.156}$$

The advantage of this representation lies in the separation of two different sources of complexity: the matrix $\mathbf{H}$ is random but independent of the function f, whereas the matrix $\mathbf{T}$ depends on f but is not random.

Furthermore, if we assume that $\mathbb{E}H_{ij} = 0$, we readily find that H_{ij} are identically distributed random variables with zero mean and variance equal to $1/N$. We will assume, in addition, that H_{ij} are independent Gaussian random variables. This assumption largely simplifies calculations but may limit applicability of our results at high densities of points ρ.

Example 6.14.1 (random Green's matrix) The purpose of this example is to study eigenvalue distributions of certain large Euclidean random matrices that appear

in problems of wave propagation in random media. In the simplest case of scalar waves the propagation is described by a scalar wave equation, so the function f that will be of interest to us is the Green's function $G\left(\mathbf{r}_i, \mathbf{r}_j\right)$ of the Helmholtz equation

$$\left(\nabla^2 + k_0^2 + i\varepsilon\right) G\left(\mathbf{r}_i, \mathbf{r}_j\right) = -\frac{4\pi}{k_0}\delta\left(\mathbf{r}_i - \mathbf{r}_j\right)$$

where ε is a positive infinitesimal. It is easy to check that

$$G\left(\mathbf{r}_i, \mathbf{r}_j\right) = \frac{\exp\left(ik_0\left|\mathbf{r}_i - \mathbf{r}_j\right|\right)}{k_0\left|\mathbf{r}_i - \mathbf{r}_j\right|}$$

A random Green's matrix is defined as

$$A_{ij} = \left(1 - \delta_{ij}\right)\frac{\exp\left(ik_0\left|\mathbf{r}_i - \mathbf{r}_j\right|\right)}{\left|\mathbf{r}_i - \mathbf{r}_j\right|} \tag{6.157}$$

where $k_0 = 2\pi/\lambda_0$ and λ_0 is the wavelength. We assume that the N points $\mathbf{r}_i$ are chosen randomly inside a three-dimensional ($d = 3$) sphere of radius R. This non-Hermitian Euclidean random matrix is of special importance in the context of wave propagation in disordered media because its elements are proportional to the Green's function of the Helmholtz equation, with $\mathbf{r}_i$, which may be thought of as positions of point-like scattering centers.

For each realization of the random matrix (6.157), its eigenvalues $\lambda_i, i = 1, \ldots, N$ obey

$$\sum_{i=1}^{N}\lambda_i = 0, \qquad \operatorname{Im}\lambda_i > -1, \quad i = 1, \ldots, N \tag{6.158}$$

Very generally, the eigenvalue density of the matrix defined by (6.157) depends on two dimensionless parameters: the number of points per wavelength cubed $\rho\lambda_0^3$ and the second moment of

$$\mathbb{E}|\lambda|^2 = \gamma = 9N/8\left(k_0R\right)^2$$

We now deal with the borderline of the support of eigenvalues, which is easier to visualize. For a low-density $\rho\lambda_0^3 \leqslant 10$, a simple equation

$$|\lambda|^2 \simeq 2\gamma\left(-8\gamma\frac{\operatorname{Im}\lambda}{3|\lambda|^2}\right) \tag{6.159}$$

gives satisfactory results. For $\gamma << 1$, the density of eigenvalues is roughly uniform within a circular domain of the radius, (see Figure 6.24(a)). The domain grows in size and shifts up with increasing γ. At $\gamma \geq 1$, it starts to "feel" the "wall" $\operatorname{Im}\lambda = -1$ and deforms Figure 6.24(b).

Another curve for the borderline of eigenvalue distribution is given by

$$|\lambda|^2 = \frac{8\gamma}{\sqrt{3}\pi}\sqrt{1 + \operatorname{Im}\lambda}\left(1 + \frac{|\lambda|^2}{|\lambda|^2 + 4\gamma}\right) \tag{6.160}$$

□

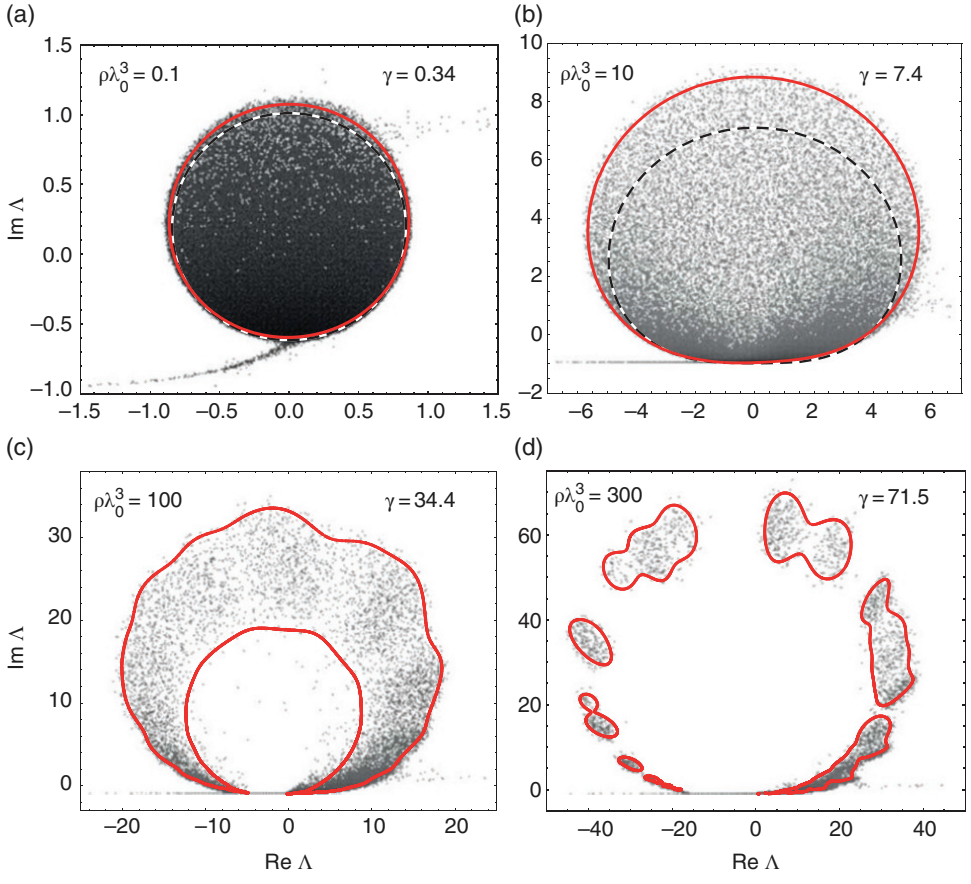


Figure 6.24 Density plots of the logarithm of eigenvalue density of the N×N random Green's matrix (6.157) obtained by numerical diagonalization of 10 realizations of the matrix for N= 10^4. The solid lines represent the borderlines of the support of eigenvalue density following from the theory. The dashed lines show the diffusion approximation (6.160). From (a) to (d), we keep increasing Y. Source: Reproduced with permission from [321].

Now let us turn to the general framework for analysis, closely following [320]. Consider a singly connected three-dimensional region of space V. Let $\{\psi_m(\mathbf{r})\}$ be an orthonormal basis in V, such that

$$\int_V d^3\mathbf{r}\psi_m(\mathbf{r})\psi_n^*(\mathbf{r}) = \delta_{mn} \tag{6.161}$$

We now show that an arbitrary $N \times N$ Euclidean random matrix $\mathbf{A}$ with elements

$$A_{ij} = f(\mathbf{r}_i, \mathbf{r}_j), \quad i,j = 1,\ldots,N \tag{6.162}$$

where f is a sufficiently well-behaved function of $\mathbf{r}_i, \mathbf{r}_j \in V$, can be represented as

$$\mathbf{A} = \mathbf{HTH}^H \tag{6.163}$$

Here $\mathbf{H}$ is an $N \times N$ matrix with elements

$$H_{im} = \sqrt{\frac{V}{N}} \psi_m \left(\mathbf{r}_i \right) \tag{6.164}$$

We use V to denote the considered three-dimensional region of space as well as its volume, and $\mathbf{T}$ an $M \times M$ matrix to be defined below. The size M of the matrix $\mathbf{T}$ can be arbitrary and, in fact, M will be infinite for the majority of functions $f\left(\mathbf{r}_i, \mathbf{r}_j \right)$.

To establish (6.163), we write the elements explicitly as

$$A_{ij} = \frac{V}{N} \sum_{m,n} T_{mn} \psi_m \left(\mathbf{r}_i \right) \psi_n^* \left(\mathbf{r}_j \right) \tag{6.165}$$

where we have used (6.164) and the definition of matrix multiplication. Multiplying this equation by $\psi_{m'}^* \left(\mathbf{r}_i \right) \psi_{n'} \left(\mathbf{r}_j \right)$, integrating over $\mathbf{r}_i$ and $\mathbf{r}_j$ and using the orthogonality of the basis functions $\psi_m \left(\mathbf{r} \right)$, we readily obtain

$$T_{mn} = \frac{V}{N} \int_V d^3 \mathbf{r}_i \int_V d^3 \mathbf{r}_j f \left(\mathbf{r}_i, \mathbf{r}_j \right) \psi_m^* \left(\mathbf{r} \right) \psi_n \left(\mathbf{r} \right) \tag{6.166}$$

When the points $\left\{ \mathbf{r}_i \right\}$ are chosen inside V randomly, $\mathbf{A}$ and $\mathbf{H}$ become *random matrices*, whereas $\mathbf{T}$ is always a nonrandom matrix independent of $\left\{ \mathbf{r}_i \right\}$ and determined uniquely by the function f, the region V, and the choice of the orthonormal basis $\left\{ \psi_m \left(\mathbf{r} \right) \right\}$. We will limit our consideration to the case when the spatial integral of any basis function $\left\{ \mathbf{r}_i \right\}$ that contributes to (6.165) vanishes[1]:

$$\int_V d^3 \mathbf{r} \psi_m \left(\mathbf{r} \right) = 0 \tag{6.167}$$

The elements H_{im} of $\mathbf{H}$ are then independent random variables having zero means and variances equal to $1/N$:

$$\begin{aligned} \mathbb{E} H_{im} &= \frac{1}{V} \int_V d^3 \mathbf{r}_i \sqrt{\frac{V}{N}} \psi_m \left(\mathbf{r}_i \right) = 0 \\ \mathbb{E} \left[H_{im} H_{jn}^* \right] &= \frac{1}{V^2} \int_V d^3 \mathbf{r}_i \int_V d^3 \mathbf{r}_j \frac{V}{N} \psi_m \left(\mathbf{r}_i \right) \psi_n^* \left(\mathbf{r}_j \right) = \mathbb{E} \left[H_{im} \right] \mathbb{E} \left[H_{jn}^* \right] = 0, \quad i \neq j \\ \mathbb{E} \left[H_{im} H_{in}^* \right] &= \frac{1}{V} \int_V d^3 \mathbf{r}_i \frac{V}{N} \psi_m \left(\mathbf{r}_i \right) \psi_n^* \left(\mathbf{r}_i \right) = \frac{1}{N} \delta_{mn} \end{aligned} \tag{6.168}$$

The representation (6.163) is very useful because it can be handled using the powerful mathematical arsenal of the so-called free random variable theory. For random matrices, the notion of asymptotic freeness [126] is equivalent to the notion of statistical independence that we are familiar with for random variables.

Three fundamental objects of the free random variable theory, defined for any Hermitian matrix $\mathbf{A}$ will be useful for us in this section: the usual Green's function

$$G(z) = \frac{1}{N} \mathbb{E} \left[\mathrm{Tr} \left(z \mathbf{I}_N - \mathbf{A} \right)^{-1} \right] \tag{6.169}$$

where $\mathbf{I}$ is an identity matrix of $N \times N$. The Blue's function is defined as the functional inverse of the Green's function $G(z)$

$$B\left[G(z) \right] = z, \tag{6.170}$$

1 This restricts the class of functions $f\left(\mathbf{r}_i, \mathbf{r}_j \right)$ to which our analysis applies but will be sufficient for us here.

and the S-transform of the probability distribution of eigenvalues defined through an auxiliary function $\chi(z)$:

$$S(z) = \frac{1+z}{z}\chi(z), \qquad \frac{1}{\chi(z)} G\left[\frac{1}{\chi(z)}\right] - 1 = z \tag{6.171}$$

If two Hermitian random matrices **A** and **B** are asymptotically free, the Blue's function $B_{\mathbf{C}}(z)$ of their sum $\mathbf{C} = A + B$ is equal to the sum of individual Blue's functions $B_{\mathbf{A}}(z)$ and $B_{\mathbf{B}}(z)$, minus $1/z$. The S-transform of the matrix product $\mathbf{C} = AB$ can be found by multiplying the individual S-transforms of **A** and **B**. Once the Blue's function or the S-transform corresponding to the random matrix **C** are found, its Green's function $G(z)$ can be calculated either from (6.170) or from (6.171). The probability density of the eigenvalues λ of **C** is then determined in the usual way:

$$p(\lambda) = -\frac{1}{\pi} \lim_{\epsilon \to \infty} \operatorname{Im} G(\lambda + i\epsilon) \tag{6.172}$$

The functions $G(z)$, $B(z)$ and $S(z)$ all contain the same full information about the statistical distribution of eigenvalues λ as $p(\lambda)$. The Green's function can be represented as an infinite series with coefficients in front of consecutive powers of $1/z$ equal to statistical moments of λ:

$$G(z) = \sum_{k=0}^{\infty} \mathbb{E}\left[\lambda^k\right] \frac{1}{z^{k+1}}.$$

We have, thus, the moments

$$\mathbb{E}\left[\lambda^k\right] = \frac{1}{(k+1)!} \frac{d^{k+1}}{d(1/z)^{k+1}}\bigg|_{z\to\infty}, \tag{6.173}$$

where z is assumed real. Using this equation and (6.170) we readily derive an expression for the k-th moment $\mathbb{E}\left[\lambda^k\right]$ in terms of Blue's function $B(z)$:

$$\mathbb{E}\left[\lambda^k\right] = \frac{1}{(k+1)!}\left[-\frac{B^2(z)}{B'(z)}\frac{d}{dz}\right]^k \left[-\frac{B^2(z)}{B'(z)}\right]\bigg|_{z\to 0}, \tag{6.174}$$

where $B'(z) = dB(z)/dz$. If we introduce the R-transform [52] $R(z) = B(z) - 1/z$, the average eigenvalue and the variance become

$$\mathbb{E}\lambda = R(0) \text{ and } \operatorname{var}\lambda = \mathbb{E}[\lambda - \mathbb{E}\lambda]^2 = R'(z)\big|_{z\to 0}$$

respectively.

For matrix **A** of the form (6.163), the free random variable theory provides a number of mathematical theorems that we will exploit in the future's research. In particular, we have [52] that

$$S_{\mathbf{A}}(z) = \frac{1}{z + M/N} S_{\mathbf{T}}\left(\frac{N}{M}z\right) \tag{6.175}$$

if **T** is a Hermitian nonnegative random matrix independent of **H** and the limits $N, M \to \infty$ are taken at a constant M/N. Using (6.175), we derive a relation between the Blues function of **A** and the Green's function of **T**:

$$B_{\mathbf{A}}(z) = \frac{1}{z}\left\{1 + \frac{M}{N}\left[\frac{1}{z}G_{\mathbf{T}}\left(\frac{1}{z}\right) - 1\right]\right\} \tag{6.176}$$

A particular case that we will consider here is when the region V is a square box of side L. A convenient set of basis functions is then given by "plane waves"

$$\psi_m(\mathbf{r}) = \frac{1}{\sqrt{V}} \exp\left(i\mathbf{q}_m \cdot \mathbf{r}\right)$$

where $\mathbf{q}_m = \left\{q_{m_x}, q_{m_y}, q_{m_z}\right\}, q_{m_x} = m_x \Delta q$ with $m_x = \pm 1, \pm 2, \ldots$ (and similarly for m_y and m_z), and $\Delta q = 2\pi/L$.

Example 6.14.2 (Eigenvalue distribution of the sinc matrix) We consider the real symmetric $N \times N$ Euclidean matrix $\mathbf{A} = S$ with elements defined through the cardinal sine (sinc) function:

$$S_{ij} = f\left(\mathbf{r}_i, \mathbf{r}_j\right) = \frac{\sin\left(k_0 \left|\mathbf{r}_i - \mathbf{r}_j\right|\right)}{k_0 \left|\mathbf{r}_i - \mathbf{r}_j\right|} \tag{6.177}$$

Here k_0 is a constant and the vector $\mathbf{r}_i$ defines positions of N randomly chosen points inside a three-dimensional cube of side L. The first important property of the matrix $\mathbf{S}$ is the positiveness of its eigenvalues: $\lambda_i(\mathbf{S}) > 0, \quad i = 1, \ldots, N$. Indeed, the Fourier transform of the function $f(\Delta\mathbf{r})$ in (6.177) is positive and hence $f(\Delta\mathbf{r})$ is a function of positive type. An Euclidean matrix defined through a function of positive type is positive definite and hence has only positive eigenvalues. The matrix $\mathbf{T}$ corresponding to $\mathbf{S}$ can be found from (6.166):

$$T_{mn} = \frac{N}{V^2} \int_V d^3\mathbf{r}_1 \int_V d^3\mathbf{r}_2 \frac{\sin\left(k_0 \left|\mathbf{r}_1 - \mathbf{r}_2\right|\right)}{k_0 \left|\mathbf{r}_1 - \mathbf{r}_2\right|} \exp\left(-i\mathbf{q}_m \cdot \mathbf{r}_1 + i\mathbf{q}_n \cdot \mathbf{r}_2\right) \tag{6.178}$$

Unfortunately, it is impossible to calculate this double integral exactly in a box. However, introducing new variables of integration $\mathbf{R} = \mathbf{r}_1 + \mathbf{r}_2$ and $\Delta\mathbf{r} = \mathbf{r}_2 - \mathbf{r}_1$ and limiting the integration over $\Delta\mathbf{r}$ to the region $|\Delta\mathbf{r}| < L/2\alpha$, with $\alpha \approx 1$ a numerical constant to be fixed later, we obtain an approximate result

$$\begin{aligned} T_{mn} &\approx \tfrac{N}{V^2} \textstyle\int_V d^3\mathbf{R} \exp\left(-i\left(\mathbf{q}_m - \mathbf{q}_n\right) \cdot \mathbf{R}\right) \\ &\quad \textstyle\int_{|\Delta\mathbf{r}|<L/2\alpha} d^3\Delta\mathbf{r} \frac{\sin(k_0\Delta r)}{k_0\Delta r} \exp\left(i\left(\mathbf{q}_m + \mathbf{q}_n\right) \cdot \Delta\mathbf{r}/2\right) \\ &= \delta_{mn} \tfrac{2\pi^2 N}{k_0 q_m} \tfrac{L}{2\alpha\pi} \left\{ \operatorname{sinc}\left[\left(q_m - k_0\right) \tfrac{L}{2\alpha}\right] - \operatorname{sinc}\left[\left(q_m + k_0\right) \tfrac{L}{2\alpha}\right] \right\} \end{aligned} \tag{6.179}$$

This expression is still too involved to be useful. The second sinc function in (6.179) is always smaller than $2\alpha/k_0 L$ (because $q_m = |\mathbf{q}_m| > 0$ and $k_0 > 0$) and hence can be dropped in the limit of large $k_0 L >> 1$, considered in this section. Besides, because the first sinc function in (6.179) is peaked around $q_m = k_0$, we replace it by a boxcar function $\prod\left[\left(q_m - k_0\right) L/2\alpha\pi\right]$, where, the boxcar function is defined as $\prod(x) = 1$ for $|x| < 1/2$ and $\prod(x) = 0$ otherwise. The coefficient in front of $(q_m - k_0)$ in the argument of $\prod$ is chosen to ensure that the integral of the latter over q_m from 0 to ∞ is equal to the same integral of the sinc function. We then obtain

$$T_{mn} \simeq \frac{2\pi^2 N}{k_0 q_m} \frac{L}{2\alpha\pi} \prod\left[\left(q_m - k_0\right) \frac{L}{2\alpha}\right] \delta_{mn}$$

which is different from zero only for $\mathbf{q}_m$s inside a spherical shell of radius k_0 and thickness $\frac{L}{2\alpha\pi}$. In addition, for all $\mathbf{q}_m$s inside the shell the value of T_{mn} is the same and equal to N/M with $M = \alpha\left(k_0 L\right)^2/\pi >> 1$, the number of inside the shell Equation (6.179) then yields

$$\mathbf{S} = \frac{N}{M}\mathbf{H}\mathbf{H}^H \tag{6.180}$$

which is equivalent to (6.163) with a $M \times M$ matrix $\mathbf{T} = \frac{N}{M}\mathbf{I}_M$. To obtain the R-transform of (6.180), we see Example 5.8.8 for the product of two i.i.d. random matrices. We then readily find

$$G_{\mathbf{T}}(z) = \frac{1}{M}\,\mathrm{Tr}\left[z\mathbf{I} - \frac{N}{M}\mathbf{I}\right]^{-1} = \frac{1}{z - N/M}$$

and from (6.176):

$$B_{\mathbf{S}}(z) = (1 - \beta z)^{-1} + 1/z$$

with $\beta = N/M$. This is the Blue's function of the famous Marchenko–Pastur law

$$p(\lambda) = \left(1 - \frac{1}{\beta}\right)^{+}\delta(\lambda) + \frac{\sqrt{(\lambda - a)(b - \lambda)}}{2\pi\beta\lambda} \tag{6.181}$$

where $a = \left(1 - \sqrt{\beta}\right)^2, b = \left(1 + \sqrt{\beta}\right)^2$ and $x^+ = \max(x, 0)$. The distribution of eigenvalues of the matrix (6.177) is therefore parameterized by a single parameter β equal to the variance of this distribution, as can be easily checked from (6.181): $\mathrm{var}(\lambda) = \beta$. Although our derivation of (6.181) was based on several approximations, the average value of λ, $\mathbb{E}\lambda = 1$, following from this equation, is exact. The second moment of λ,

$$\mathbb{E}\lambda^2 = \frac{1}{N}\mathbb{E}\left[\mathrm{Tr}\,\mathbf{S}\right] = 1 + \frac{aN}{\left(k_0 L\right)^2} \tag{6.182}$$

where the numerical constant a is given by

$$a = \frac{1}{2}\int_{\text{unit cube}} d^3\mathbf{u}_1 \int_{\text{unit cube}} d^3\mathbf{u}_2 \frac{1}{\left|\mathbf{u}_1 - \mathbf{u}_2\right|^2} \simeq 2.8$$

with the integrations running over the volume of a cube of unit side. By requiring that the second moment $1 + \beta$ of the distribution (6.181) coincides with (6.182), we can now fix the value of α that remained arbitrary until now. We obtain $\alpha = \pi/a \simeq 1.12$ and

$$\beta = \frac{2.8N}{\left(k_0 L\right)^2}$$

□

Example 6.14.3 (Eigenvalue distribution of cosc matrix) Let us now consider a Euclidean random matrix with elements defined using the cardinal cosine (cosc) function:

$$C_{ij} = \left(1 - \delta_{ij}\right)\cos\left(k_0\left|\mathbf{r}_i - \mathbf{r}_j\right|\right)/k_0\left|\mathbf{r}_i - \mathbf{r}_j\right|, \quad i, j = 1, \ldots, N \tag{6.183}$$

The prefactor $1-\delta_{ij}$ allows us to deal with the divergence of the function $\cos(x)/x$ for $x \to 0$. Proceeding as in the previous example, we find

$$T_{mn} \simeq \frac{4\pi N}{k_0 V} \frac{1}{q_m^2 - k_0^2} \delta_{mn} \tag{6.184}$$

under the same approximations as in (6.179). The matrix $\mathbf{T}$ defined in (6.184) has infinite size.

For details of this example, we refer to [320]. □

Let us recall some results from free probability that will be needed in the next example.

The extension of free probability theory and, in particular, the generalization of the concept of the Blue's function, to non-Hermitian matrices is natural in quaternion space [322, 323]. The 2×2 matrix $\mathbf{Q}$ is an arbitrary quaternion in the matrix representation

$$\mathbf{Q} = \begin{pmatrix} a & ib^* \\ b & a^* \end{pmatrix} \tag{6.185}$$

For an arbitrary Q defined above, we can use algebraic properties of quaternions to show that the following addition law holds [322, 323]

$$R_{\mathbf{X}_1+\mathbf{X}_2}(\mathbf{Q}) = R_{\mathbf{X}_1}(\mathbf{Q}) + R_{\mathbf{X}_2}(\mathbf{Q}) \tag{6.186}$$

where $\mathbf{X}_1$ and $\mathbf{X}_2$ are two *non-Hermitian*, asymptotically free random matrices.

Now consider the non-Hermitian complex matrix $\mathbf{X}_1 + i\mathbf{X}_2$, $\mathbf{X}_1$ and $\mathbf{X}_2$ are two asymptotically free *Hermitian* matrices with known R-transforms. Jarosz and Nowak (2004, 2006) [322, 323] showed that the problem reduces to solving a simple system of three equations with three unknown variables, complex u, v, and real t:

$$\begin{aligned} R_{\mathbf{X}_1}(u) &= x + \frac{t-1}{u} \\ R_{\mathbf{X}_2}(v) &= y - \frac{t}{v} \\ |u| &= |v| \end{aligned} \tag{6.187}$$

where $z = x + iy$. We express u and v via t from the first two equations, substitute the results into the third equation, and then solve for t. The resolvent and the correlator are then given by

$$g_{\mathbf{X}_1+i\mathbf{X}_2}(z) = \operatorname{Re} u - i \operatorname{Re} v \tag{6.188}$$

$$c_{\mathbf{X}_1+i\mathbf{X}_2}(z) = (\operatorname{Re} u)^2 + (\operatorname{Re} v)^2 - |u|^2 \tag{6.189}$$

Equation for the borderline $z \in \delta D$ of the eigenvalue domain follows from $c_{\mathbf{X}_1+i\mathbf{X}_2}(z) = 0$.

Example 6.14.4 (Eigenvalue distribution of the cosc $+i$ sinc **matrix and the complex expc matrix)** The matrices $\mathbf{C}$ and $\mathbf{S}$ can be combined in a single complex non-Hermitian matrix: $\mathbf{C} + i(\mathbf{S} - \mathbf{I})$. The theory of free random variables allows one to study the statistical distribution of the complex eigenvalues of this matrix based on the properties of the matrices $\mathbf{C}$ and $\mathbf{S}$ that we considered in the previous examples. This, however, requires asymptotic freeness of $\mathbf{C}$ and $\mathbf{S}$. The matrices $\mathbf{C}$ and $\mathbf{S}$ defined

by (6.177) and (6.183) through the same set of points $\mathbf{r}_i$ turn out to be not asymptotically free. We, therefore, start our study of non-Hermitian Euclidean random matrices by the case of a matrix

$$\mathbf{X} = \mathbf{C} + i\left(\mathbf{S}' - \mathbf{I}\right) \tag{6.190}$$

where two *different and independent* sets of points $\mathbf{r}_i$ and $\mathbf{r}_i^{'}$ are used to define the real and imaginary parts of $\mathbf{X}$:

$$S_{ij} = \left(1-\delta_{ij}\right)\frac{\sin\left(k_0\left|\mathbf{r}_i-\mathbf{r}_j\right|\right)}{k_0\left|\mathbf{r}_i-\mathbf{r}_j\right|}, \qquad C_{ij} = \left(1-\delta_{ij}\right)\frac{\cos\left(k_0\left|\mathbf{r}_i-\mathbf{r}_j\right|\right)}{k_0\left|\mathbf{r}_i-\mathbf{r}_j\right|} \tag{6.191}$$

Since $\mathbf{X}$ is of the form $\mathbf{X}_1 + i\mathbf{B}_2$, where $\mathbf{X}_1 = \mathbf{C}$ and $\mathbf{X}_2 = \mathbf{S}' - \mathbf{I}_N$ are two asymptotically free Hermitian matrices, we can make use of (6.187), (6.188) and (6.189) to calculate the resolvent $g(z)$ and the eigenvector correlator $c(z)$ of $\mathbf{X}$. In the limit of $\gamma << 1$, the R-transforms of $\mathbf{X}_1$ and $\mathbf{X}_2$ are those of Gaussian and Wishart matrices, respectively:

$$g\left(z=x+iy\right) = \frac{x}{2\gamma} - \frac{i}{2}\left[\frac{y}{\gamma\left(1+y\right)} + \frac{1}{2+y}\right] \tag{6.192}$$

$$c\left(z=x+iy\right) = \left(\frac{x}{2\gamma}\right)^2 + \frac{1}{4}\left[\frac{y}{\gamma\left(1+y\right)} - \frac{1}{2+y}\right]^2 - \frac{1}{\gamma\left(1+y\right)\left(2+y\right)} \tag{6.193}$$

The correlator (6.193) must vanish on the borderline δD of the eigenvalue domain. We therefore readily obtain equation for the borderline on the complex plane:

$$x^2 + \left(\frac{y}{\left(1+y\right)} - \frac{\gamma}{2+y}\right)^2 - \frac{4\gamma}{\left(1+y\right)\left(2+y\right)} = 0 \tag{6.194}$$

The probability density inside the domain delimited by (6.194) is

$$\begin{aligned} p\left(x,y\right) &= \tfrac{1}{2\pi}\left[\partial_x \operatorname{Re} g\left(z\right) - \partial_x \operatorname{Re} g\left(z\right)\right] \\ &= \tfrac{1}{2\pi}\left[\tfrac{1}{\gamma} + \tfrac{1}{\gamma\left(1+y\right)^2} - \tfrac{1}{\left(2+y\right)^2}\right] \end{aligned} \tag{6.195}$$

By analogy with the cardinal sine and cosine functions, a "cardinal complex exponent" function can be defined as $f\left(x\right) = \exp\left(ix\right)/x$. The Euclidean random matrix $\mathbf{G}$ corresponding to this function has elements

$$G_{ij} = f\left(\mathbf{r}_i - \mathbf{r}_j\right) = \left(1-\delta_{ij}\right)\frac{\exp\left(ik_0\left|\mathbf{r}_i-\mathbf{r}_j\right|\right)}{k_0\left|\mathbf{r}_i-\mathbf{r}_j\right|} \tag{6.196}$$

This matrix has a particular importance in the problem of wave scattering by an ensemble of N point-like scatterers. Although the matrix $\mathbf{G}$ is similar to the matrix $\mathbf{X}$, the analytic study of its properties is much more involved. See [321] for an analytical theory. □

6.15 Random Matrices with Independent Entries and the Circular Law

In this section we consider two ensembles of random matrices with **independent** entries. Before we state the circular law, we first define a class of Hermitian random matrices with independent entries originally introduced by Wigner (1958) [109].

Definition 6.15.1 (Wigner random matrices) Let ξ be a complex random variable with mean zero and unit variance, and let ς be a real random variable with mean zero and finite variance. We say $\mathbf{X}_n$ is a Wigner matrix of size n with atom variables ξ, ς if $\mathbf{X}_n = \left(X_{ij}\right)_{i,j=1}^{n}$ is a random Hermitian $n \times n$ matrix that satisfies the following conditions:

- Independent random variables. $\left\{X_{ij} : 1 \leqslant i \leqslant j \leqslant n\right\}$ is a collection of independent random variables.
- Entries above the diagonal ones. $\left\{X_{ij} : 1 < i < j \leqslant n\right\}$ is a collection of independent and identically distributed (i.i.d.) copies of ξ.
- Diagonal entries. $\left\{X_{ii} : 1 \leqslant i \leqslant n\right\}$ is a collection of i.i.d. copies of ς.

The prototypical example of a Wigner real symmetric matrix is the Gaussian orthogonal ensemble (GOE). The GOE is defined by the probability distribution

$$\mathbb{P}\left(d\mathbf{M}\right) = \frac{1}{Z_n^{(\beta)}} \exp\left(-\frac{\beta}{4} \operatorname{Tr} \mathbf{M}^2\right) d\mathbf{M} \tag{6.197}$$

on the space of $n \times n$ real symmetric matrices when $\beta = 1$ and $d\mathbf{M}$ refers to the Lebesgue measure on the $n(n+1)/2$ different elements of the matrix. Here $Z_n^{(\beta)}$ denotes the normalization constant. So for a matrix $\mathbf{X}_n = \left(X_{ij}\right)_{i,j=1}^{n}$ drawn from the GOE, the elements $\left\{X_{ij} : 1 \leqslant i \leqslant j \leqslant n\right\}$ are independent Gaussian random variables with mean zero and variance $1 + \delta_{ij}$. The classical example of a Wigner Hermitian matrix is the Gaussian unitary ensemble (GUE). The GUE is defined by the probability distribution given in (6.197) with $\beta = 2$, but on the space of $n \times n$ Hermitian matrices. Thus, for a matrix $\mathbf{X}_n = \left(X_{ij}\right)_{i,j=1}^{n}$ drawn from the GUE, the n^2 different real elements of the matrix

$$\left\{\operatorname{Re}\left(Y_{ij}\right) : 1 \leqslant i \leqslant j \leqslant n\right\} \cup \left\{\operatorname{Im}\left(Y_{ij}\right) : 1 \leqslant i \leqslant j \leqslant n\right\}$$

are independent Gaussian random variables with mean zero and variance $(1 + \delta_{ij})/2$. A classical result for Wigner random matrices is Wigner's semicircle law Wigner (1958) [109, Theorem 2.5].

Theorem 6.15.2 (Wigner's semicircle law) Let ξ be a complex random variable with mean zero and unit variance, and let ς be a real random variable with mean zero and finite variance. For each $n \geq 1$, let $\mathbf{X}_n$ be a Wigner matrix of size n with atom variables ξ, ς, and let $\mathbf{A}_n$ be a deterministic $n \times n$ Hermitian matrix with rank $o(n)$. Then the empirical spectral distribution of $\frac{1}{\sqrt{n}}\left(\mathbf{X}_n + \mathbf{A}_n\right)$ converges almost surely to the semicircle distribution $F_{sc}(x)$ as $n \to \infty$, where

$$F_{sc}\left(x\right) = \int_{-\infty}^{x} f_{sc}\left(t\right) dt, \qquad f_{sc}\left(x\right) = \begin{cases} \frac{1}{2\pi}\sqrt{4 - x^2}, & \text{if } |x| \leqslant 2 \\ 0, & \text{if } |x| > 2 \end{cases}$$

Wigner's semicircle law holds in the case where the entries of $\mathbf{X}_n$ are not identically distributed (but are still independent) provided the entries satisfy a Lindeberg-type condition. See [163, Theorem 2.9] for further details.

Now we are ready for the circular law. The circular law is another milestone in the development of random-matrix theory. The circular-law theorem states that the empirical spectral distribution of an $n \times n$ random matrix with i.i.d. entries of variance $1/n$ tends to the uniform law on the unit disc of the complex plane as the dimension n tends to infinity. This phenomenon is the non-Hermitian counterpart of the semi circular limit for Wigner random Hermitian matrices, and the quarter-circular limit for Marchenko-Pastur random covariance matrices

We now consider an ensemble of random matrices with i.i.d. entries. That is, we consider a random $n \times n$ matrix $\mathbf{X}_n$ whose entries are i.i.d. copies of a random variable ξ. In this case, we say $\mathbf{X}_n$ is an *i.i.d. random matrix*, and we refer to ξ as the atom variable of $\mathbf{X}_n$. When ξ is a standard complex Gaussian random variable, $\mathbf{X}_n$ can be viewed as a random matrix drawn from the probability distribution

$$\mathbb{P}(d\mathbf{M}) = \frac{1}{\pi^{n^2}} \exp\left(-\operatorname{Tr}\left(\mathbf{M}\mathbf{M}^H\right)\right) d\mathbf{M}$$

on the set of complex $n \times n$ matrices. Here $d\mathbf{M}$ denotes the Lebesgue measure on the $2n^2$ real entries of $\mathbf{M}$. This is known as the complex Ginibre ensemble. The real Ginibre ensemble is defined analogously. Following Ginibre (1965) [111], one may compute the joint density of the eigenvalues of a random $n \times n$ matrix $\mathbf{X}_n$ drawn from the complex Ginibre ensemble.

Mehta [103, 324] used the joint density function obtained by Ginibre to compute the limiting spectral measure of the complex Ginibre ensemble. In particular, he showed that if $\mathbf{X}_n$ is drawn from the complex Ginibre ensemble, then the ESD of $\frac{1}{\sqrt{n}}\mathbf{X}_n$ converges to the *circular law* $F_{\text{circle}}(x, y)$, where

$$F_{\text{circle}}(x, y) = \mu_{\text{circle}}\left(\left\{z \in \mathbb{C} : \operatorname{Re}(z) \leqslant x, \operatorname{Im}(z) \leqslant y\right\}\right)$$

and μ_{circle} is the uniform probability measure on the unit disk in the complex plane. Edelman (1997) [113] verified the same limiting distribution for the real Ginibre ensemble.

For the general (non-Gaussian) case, there is no formula for the joint distribution of the eigenvalues and the problem appears much more difficult. The universality phenomenon in random matrix theory asserts that the spectral behavior of a random matrix does not depend on the distribution of the atom variable ξ in the limit $n \to \infty$. In other words, one expects that the circular law describes the limiting ESD of a large class of random matrices (not just Gaussian matrices).

Let X_{ij}, $1 \leqslant i \leqslant j < \infty$, be an array of independent random variables with $\mathbb{E}X_{ij} = 0$. We consider the random matrix

$$\mathbf{X}_n = \left\{X_{ij}\right\}_{i,j=1}^{n}$$

Denote by $\lambda_1, \ldots, \lambda_n$ the eigenvalues of the matrix $\mathbf{X}_n$ and define its spectral distribution function $F_{\mathbf{A}_n}(x, y)$ by (3.6).

We say that the Circular law holds if $F_{\mathbf{A}_n}(x, y)$ converges to the distribution function $F(x, y)$ of the uniform distribution in the unit disc in $\mathbb{R}^2$. $F(x, y)$ is called the circular law.

For matrices with independent identically distributed complex normal entries, the circular law was proved by Mehta, see [103]. Almost certain convergence of $F_{\mathbf{X}_n}(x,y)$ to the circular law $F(x,y)$ under the assumption of a finite fourth $(2+\varepsilon)$ and finally of the second moment was established in [325] by Pan, Zhou and by Tao, Vu in [326, 327], respectively. The excellent tutorial by Bordenave and Chafaï is [328, 329].

Theorem 6.15.3 (Tao-Vu (2010) [55]) Let ξ be a complex random variable with mean zero and unit variance. For each $n \geq 1$, let $\mathbf{X}_n$ be a $n \times n$ matrix whose entries are i.i.d. copies of ξ, and let $\mathbf{A}_n$ be an $n \times n$ deterministic matrix. If

$$\text{rank}\left(\mathbf{A}_n\right) = o(n) \quad \text{and} \quad \sup_{n \geqslant 1} \frac{1}{n^2} \left\|\mathbf{A}_n\right\|_F^2 < \infty$$

then the ESD of $\frac{1}{\sqrt{n}}\left(\mathbf{X}_n + \mathbf{A}_n\right)$ converges almost certainly to the circular law $F_{\text{circle}}\left(x,y\right)$ as $n \to \infty$.

6.16 The Circular Law and Outliers

The random matrix $\frac{1}{\sqrt{n}}\mathbf{X}$ is perturbed by an deterministic matrix $\mathbf{A}$ such that the eigenvalues of $\frac{1}{\sqrt{n}}\mathbf{X} + \mathbf{A}$ are considered. We define the signal-to-noise ratio (SNR) in dB as

$$\text{SNR} = 10 * \log_{10}\left(\frac{\text{Tr}\left(\mathbf{A}\mathbf{A}^H\right)}{\text{Tr}\left(\mathbf{X}\mathbf{X}^H/n\right)}\right)$$

Figure 6.25, Figure 6.26 and Figure 6.27 are plotted for $n = 50$, $n = 200$, and $n = 1,000$, respectively. For other related simulation results, see Figure 6.28–6.31. From these results, we illustrate the statistical properties of outliers. We can clearly identify every corresponding eigenvalue location on the complex plane. For $n = 1,000$ in Figure 6.27, SNR is -12.2 (dB). We can consider a more general model

$$\mathbf{Z} = \frac{1}{\sqrt{n}}\mathbf{X} + \mu\mathbf{Y} \tag{6.198}$$

where $\mathbf{Y}$ is a random matrix (independent of $\mathbf{X}$) such that the columns of $\mathbf{Y}$ are i.i.d., with each column equal to $\sqrt{\frac{p}{1-p}}\phi_n$ with probability p and $-\sqrt{\frac{1-p}{p}}\phi_n$ with probability $1-p$, for some fixed $0 < p < 1$.

Code 1: I.I.D. Random Matrix perturbed by a Diagonal Matrix

```
%************************************************************
% The code is devloped by Robert C. Qiu
%
% based on the paper of Terrence Tao cited below
%
% Terrence Tao, Outliers in the spectrum of i.i.d. matrices
with bounded rank
% perturbations, Probability Theory and Related Fields,
```

```
Vol. 155, No. 1-2,
% pp. 231-263, 2013.
%*********************************************************
clear all;
    n=50;
    N_Try=5

A=zeros(n,n);

A(1,1)=2+i; A(2,2)=3;A(3,3)=2; % determinstic matrix of
low rank
A(4,4)=-2-i; A(5,5)=-1.5;A(6,6)=-2; % determinstic matrix of
low rank
A(7,7)=2-i; A(8,8)=-2+2*i;A(9,9)=-2-i*3; % determinstic
matrix of low rank
A(10,10)=-2+i; % determinstic matrix of low rank

 %A=zeros(n,n);
%A(:,1)=(2*randn(n,1)+j*randn(n,1));
```

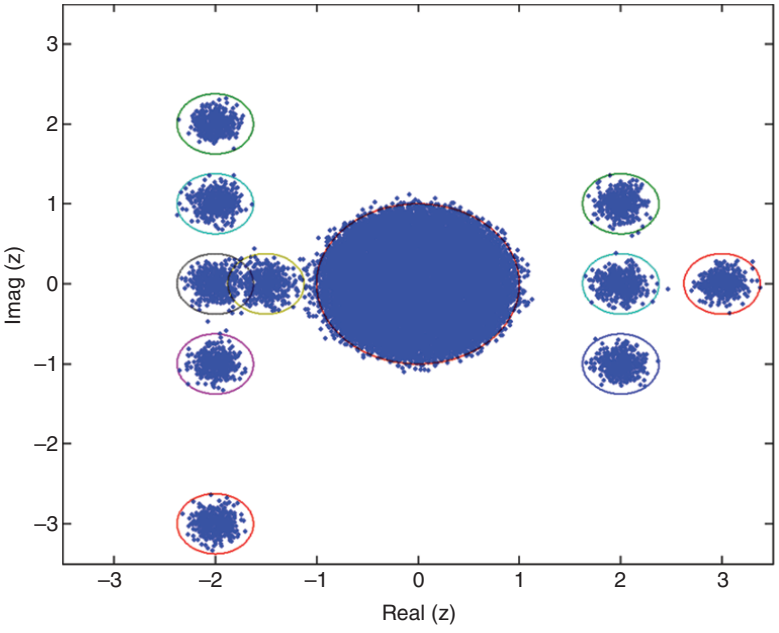


Figure 6.25 This figure shows the eigenvalues of a single $n \times n$ i.i.d. random matrix with atom distribution X defined by a white Gaussian random variable with zero mean and variance one; the eigenvalues were perturbed by adding the diagonal matrix with ten diagonal entries: $2 + i$; 3; 2; $-2 - i$; -1.5; -2; $2 - i$; $-2 + 2i$; $-2 - 3i$; $-2 + i$, corresponding to ten locations on the complex z plane. The small circles are centered at these ten locations on the complex plane, and each has a radius $n^{-\frac{1}{4}}$ where $n = 50$. Five hundred Monte Carlo trials are performed to see how stable these eigenvalues locations are. We can clearly identify every corresponding eigenvalue location on the complex plane.

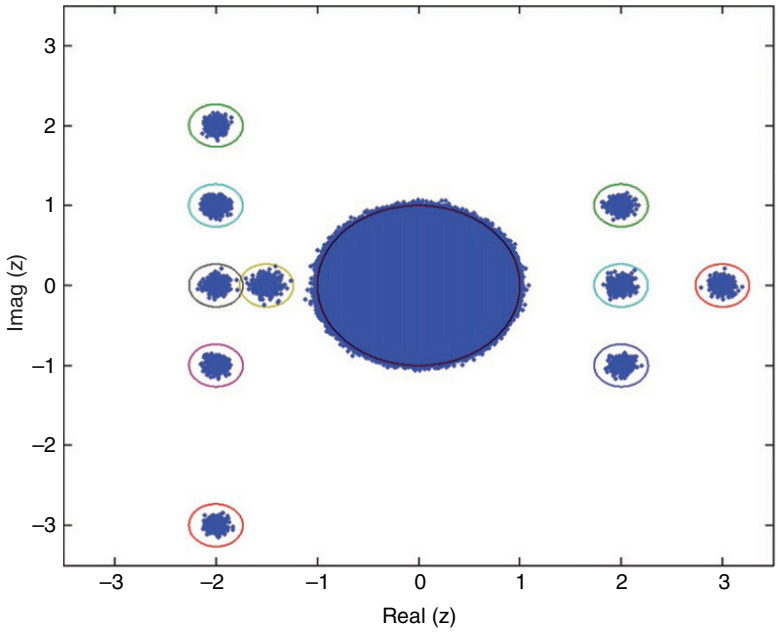


Figure 6.26 Parameters are same as Figure 6.25, except for $n = 200$.

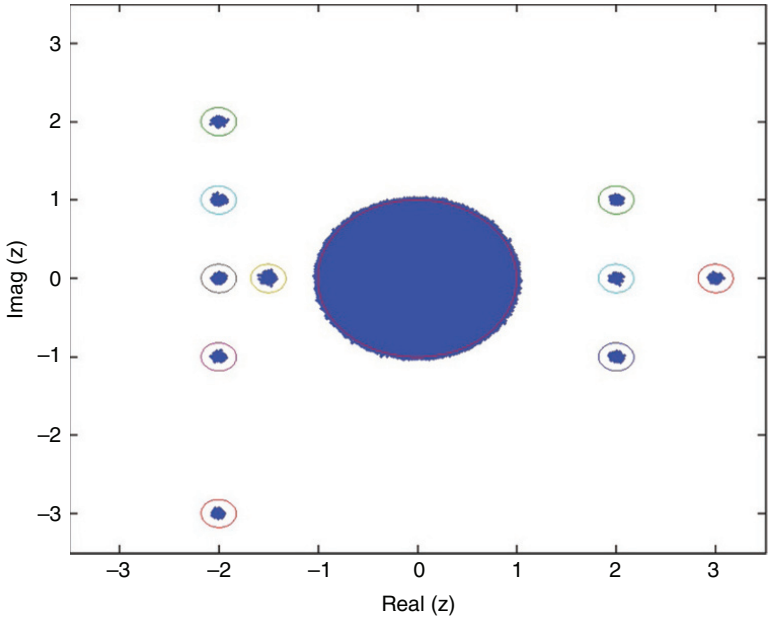


Figure 6.27 Parameters are same as Figure 6.25, except for $n = 1000$.

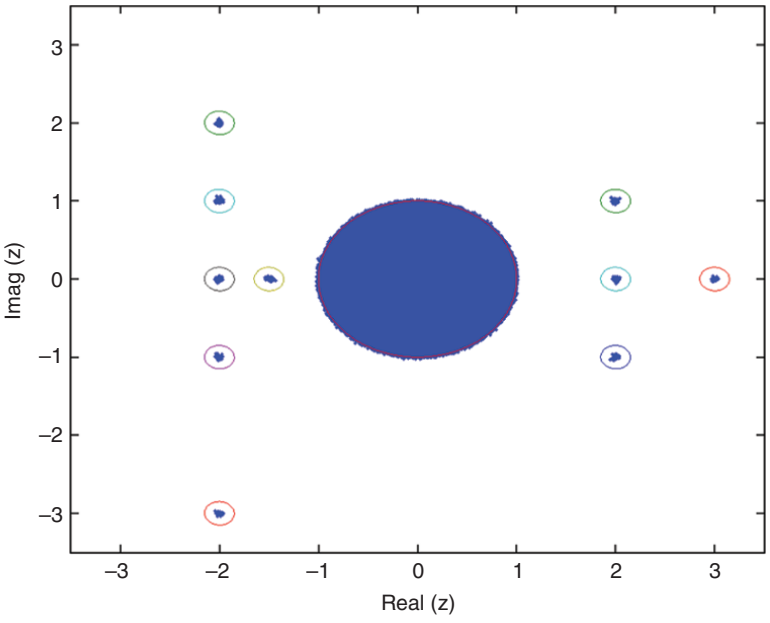


Figure 6.28 Parameters are same as Figure 6.25, except for $n = 2000$. Only 50 Monte Carlo trials are performed here, rather than 500.

```
%A(:,2)=(4*randn(n,1)+3*j*randn(n,1));

for i=1:N_Try
    X=zeros(n,n);
X=(randn(n,n)+j*randn(n,n))/sqrt(2);
% i.i.d. random matrix with i.i.d. (Gaussian)
 % complex entries
lamda=eig(X/sqrt(n)+A); % eigenvalues are complex numbers
SNR_dB=10*log10(trace(A*A')/trace(X*X'))
%*********** Figures *********
IFIG=0;
IFIG=IFIG+1;figure(IFIG);
t=0:2*pi/1000:2*pi;x=sin(t);y=cos(t); % unit circle
r=n^(-1/4); % radius of circle
plot(real(lamda),imag(lamda),'.',x,y,'r-',2+r*x,1+r*y,
3+r*x,r*y,2+r*x,r*y,...
-2+r*x,-1+r*y,-1.5+r*x,n^(-1/4)*y,-2+r*x,r*y,
2+r*x,-1+r*y,-2+r*x,2+r*y,... -2+r*x,-3+r*y,
-2+r*x,1+r*y);hold on;
axis([-3.5 3.5 -3.5 3.5])
xlabel('real(z)')
ylabel('imag(z)')
end % N_Try
```

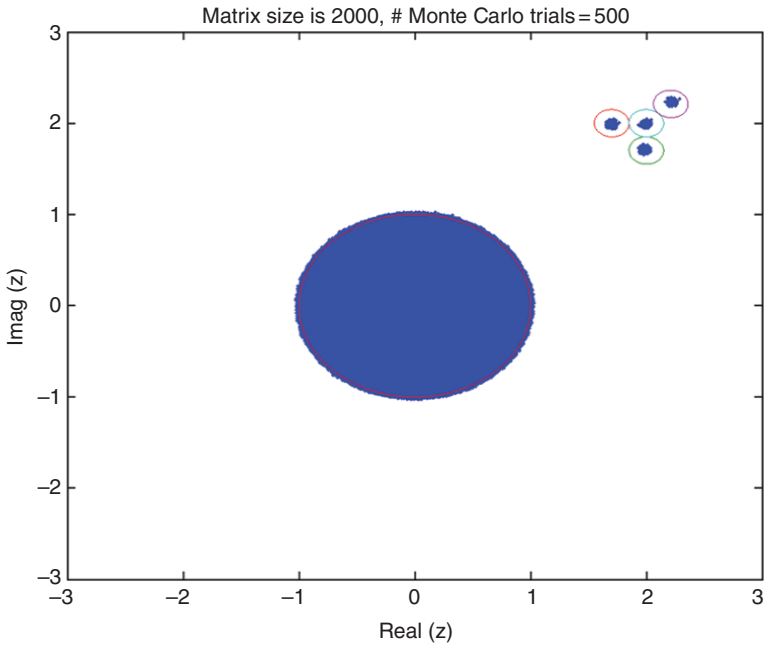


Figure 6.29 This figure shows the eigenvalues of a single $n \times n$ i.i.d. random matrix with atom distribution X defined by a white Gaussian random variable with zero mean and variance one; the eigenvalues were perturbed by adding a deterministic matrix with four eigenvalues: 2+2j;2-δ+2j;2+2j-j-δ;2+2j+$\delta/\sqrt{2}$+j$\delta/\sqrt{2}$ (their corresponding eigenvectors are random Gaussian vectors). Here $\delta = 2n^{-\frac{1}{4}}$ is the minimum distance between two eigenvalues. The small circles are centered at these four eigenvalue locations on the complex plane, respectively, and each has a radius $n^{-\frac{1}{4}}$ where $n = 2000$. Five hundred Monte Carlo trials are performed to see how stable these eigenvalues locations are. We can clearly identify every corresponding eigenvalue location on the complex plane.

```
hold off;
a=eig(A);
a(1:15)
```

Code 2: I.I.D. Random Matrix Perturbed by a Deterministic Matrix

```
%************************************************************
%
% Outliers in the spectrum of i.i.d. matrices with bounded
rank perturbations
%
%       Terrence Tao
%
% Probability Theory and Related Fields, Vol. 155, No. 1-2,
pp. 231-263, 2013.
```

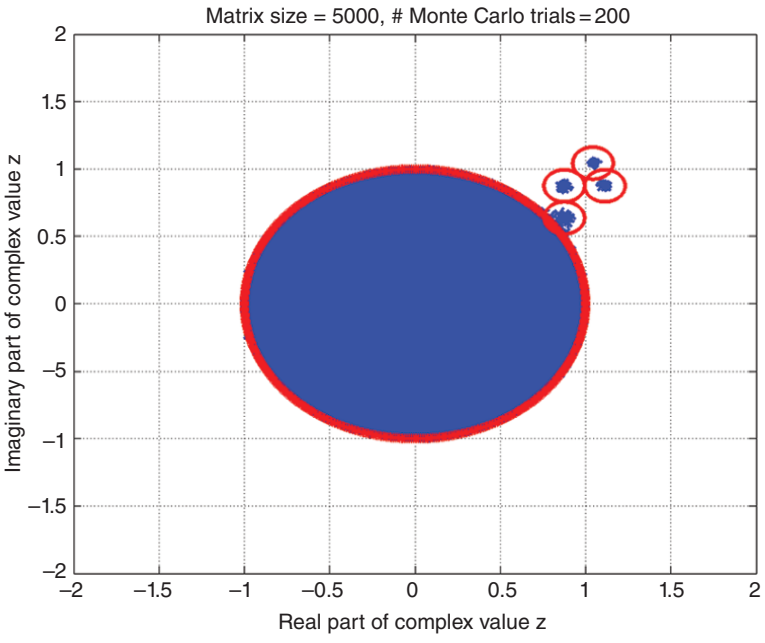


Figure 6.30 This figure shows the eigenvalues of a single $n \times n$ i.i.d. random matrix with atom distribution X defined by a white Gaussian random variable with zero mean and variance one; the eigenvalues were perturbed by adding a deterministic matrix with four eigenvalues: a+jb;a-δ+jb;a+jb-jδ;a+jb+$\delta/\sqrt{2}$+j$\delta/\sqrt{2}$ (their corresponding eigenvectors are random Gaussian vectors). Here $\delta = 2n^{-\frac{1}{4}}$ is the minimum distance between two eigenvalues, and $a = (1+\delta)/\sqrt{2}; b = (1+\delta)/\sqrt{2}$. The small circles are centered at these 4 eigenvalue locations on the complex plane, and each has a radius of $n^{-\frac{1}{4}}$. In this case $n = 5,000$. 200 Monte Carlo trials are performed to see how stable these eigenvalues locations are. We can clearly identify every corresponding eigenvalue location on the complex plane.

```
%************************************************************
clear all;
n=5000;
N_Try=5
Axis_Length=3;
IFIG=0;

A=zeros(n,n);

x1=randn(n,1); x2=randn(n,1);x3=randn(n,1);x4=randn(n,1);
x1=x1/sqrt(x1'*x1);x2=x2/sqrt(x2'*x2);x3=x3/sqrt(x3'*x3);
x4=x4/sqrt(x4'*x4);
A=(x1*x1'+j*x2*x2'+(1+j)*x3*x3'+(-1-j)*x4*x4')*2;
% matrix with rank =4


lamda=eig(A); % eigenvalues are complex numbers
```

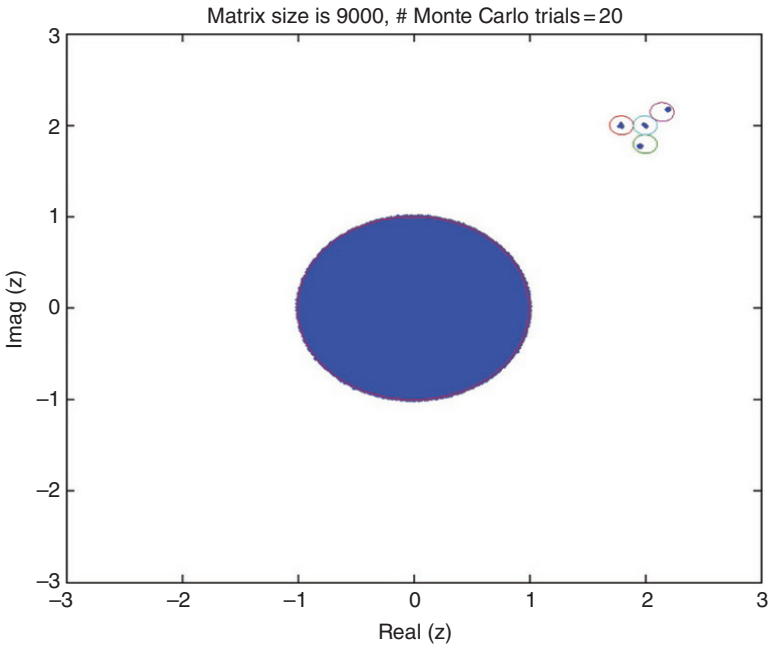


Figure 6.31 This figure shows the eigenvalues of a single $n \times n$ i.i.d. random matrix with atom distribution X defined by a white Gaussian random variable with zero mean and variance one; the eigenvalues were perturbed by adding a deterministic matrix with four eigenvalues: 2; 2+2j;2j;-2-2j (their corresponding eigenvectors are random Gaussian vectors). The small circles are centered at these four eigenvalue locations on the complex plane, and each has a radius $n^{-\frac{1}{4}}$ where $n = 9000$. Twenty Monte Carlo trials are performed to see how stable these eigenvalues locations are. We can clearly identify every corresponding eigenvalue location on the complex plane.

```
IFIG=IFIG+1;figure(IFIG);
t=0:2*pi/1000:2*pi;x=sin(t);y=cos(t); % unit circle
plot(real(lamda),imag(lamda),'*',x,y,'r-');hold on;
axis([Axis_Length*(-1) Axis_Length Axis_Length*(-1)
Axis_Length])
grid;
xlabel('real(z)')
ylabel('imag(z)')
hold off;

for i=1:N_Try X=zeros(n,n);
X=(randn(n,n)+j*randn(n,n))/sqrt(2); % i.i.d. random matrix
with i.i.d.
                                % (Gaussian) complex entries
lamda=eig(X/sqrt(n)+A); % eigenvalues are complex numbers
SNR_dB=10*log10(trace(A*A')/trace(X*X'/n))
%*********** Figures *********
IFIG=1; IFIG=IFIG+1;figure(IFIG);
```

```
t=0:2*pi/1000:2*pi;x=sin(t);y=cos(t); % unit circle
r=n^(-1/4); % radius of circle
plot(real(lamda),imag(lamda),'.',x,y,'r-',...
    2+r*x,0+r*y,2+r*x,2+r*y,0+r*x,2+r*y,-2+r*x,
    -2+r*y);hold on;
axis([Axis_Length*(-1) Axis_Length Axis_Length*(-1)
Axis_Length])
grid;
xlabel('real(z)')
ylabel('imag(z)')
end % N_Try

D=eig(A)
hold off;
```

Example 6.16.1 (orthogonal frequency-division multiplexing (OFDM) systems) Channel estimation is critical to orthogonal frequency-division multiplexing (OFDM) systems [330–332]. In this example, we follow the system model of [332] below. Our goal is to reformulate the system in terms of large random matrices and study outliers of the perturbed random matrices in this context.

We assume that the use of a cyclic prefix (CP) both preserves the orthogonality of the tones and eliminates intersymbol interference (ISI) between consecutive OFDM symbols. Further, the channel is assumed to be slowly fading, so it is considered to be constant during one OFDM symbol. The number of tones in the system is N and the length of the CP is L samples.

Under these assumptions we can regard the system as a set of *parallel* Gaussian channels, with correlated attenuation h_k. The attenuation on each tone are given by

$$h_k = G\left(\frac{k}{NT_s}\right), \quad k = 0, \ldots, N-1$$

where $G(\cdot)$ is the frequency response of the channel $g(\tau)$ during the OFDM symbol and T_s is the sampling period of the system. In matrix notation, we describe the OFDM system as

$$\mathbf{y} = \mathbf{Z}\mathbf{h} + \mathbf{x} \tag{6.199}$$

where where $\mathbf{y} = \left[y_0, \ldots, y_{N-1}\right]^T$ is the received vector

$$\mathbf{Z} = \begin{pmatrix} z_0 & & 0 \\ & \ddots & \\ 0 & & z_{N-1} \end{pmatrix}$$

is a diagonal matrix containing the transmitted signaling points, $\mathbf{h} = \left[h_0, \ldots, h_{N-1}\right]^T$ is a channel attenuation vector, and $\mathbf{x} = \left[x_0, \ldots, x_{N-1}\right]^T$ is a vector of independent identically distributed (i.i.d.) complex zero-mean Gaussian noise with variance σ^2. The random noise vector $\mathbf{x}$ is assumed to be uncorrelated with the random channel vector $\mathbf{h}$. In multiuser systems, the interference from other uses can be also modeled in the vector $\mathbf{x}$.

Putting $\mathbf{a} = \mathbf{Zh} \in \mathbb{C}^N$ within the duration of each symbol consisting of N tones, we rewrite (6.199) in terms of our standard random vector form

$$\mathbf{y} = \mathbf{x} + \mathbf{a}$$

where the random vector $\mathbf{a}$ is independent of the noise random vector $\mathbf{x}$. Assume that we consecutively use the channel for n times, representing the duration of n symbols. Then, we have

$$\mathbf{y}_i = \mathbf{a}_i + \mathbf{x}_i, \quad i = 1, \ldots, n \tag{6.200}$$

whose i-th realization of the channel may or may not independent of j-th realization of the channel, for $i \neq j, i, j = 0, \ldots, N-1$. Writing

$$\mathbf{Y} = \frac{1}{\sqrt{n}} \left(\mathbf{y}_1, \ldots, \mathbf{y}_n\right) \in \mathbb{C}^{N \times n}, \mathbf{A} = \frac{1}{\sqrt{n}} \left(\mathbf{a}_1, \ldots, \mathbf{a}_n\right) \in \mathbb{C}^{N \times n}, \mathbf{X} = \left(\mathbf{x}_1, \ldots, \mathbf{x}_n\right) \in \mathbb{C}^{N \times n}$$

we have the random matrices

$$\mathbf{Y} = \frac{1}{\sqrt{n}} \mathbf{X} + \mathbf{A} \tag{6.201}$$

where the signal random matrix $\mathbf{A}$ is independent of the noise random matrix $\mathbf{X}$. (6.201) is called the standard random matrix form of the signal-plus-noise model. We are interested in the asymptotic regime

$$n \to \infty, N \to \infty, \text{ but } n/N \to c \in (0, \infty)$$

so n and N are large but comparable, to exploit the recent developments of large dimensional random matrices in the statistics literature.

In (6.201), when the random matrix $\frac{1}{\sqrt{n}}\mathbf{X}$ is a matrix with independent identically distributed complex normal entries, the circular law for $\frac{1}{\sqrt{n}}\mathbf{X}$ has been proven. The random matrix $\frac{1}{\sqrt{n}}\mathbf{X}$ is perturbed by a deterministic (or random) matrix $\mathbf{A}$ such that the eigenvalues of $\frac{1}{\sqrt{n}}\mathbf{X} + \mathbf{A}$ exhibit outliers in the complex plane outside the unit circle. See previous figures in this section for illustrations. This outlier model was first studied by Tao (2011) [333] and was also treated previously in this section.

Now we consider the use of outliers for channel estimation or symbols decision making (demodulation). From the K outliers of the eigenvalues

$$\lambda_i \left(\frac{1}{\sqrt{n}} \mathbf{X} + \mathbf{A} \right), i = 1, \ldots, n$$

we can detect the eigenvalues

$$\lambda_k \left(\mathbf{A}\right), k = 1, \ldots, K$$

which are in turn linked with the channel vector $\mathbf{h} = \left[h_0, \ldots, h_{N-1}\right]^T$ and the transmitting signal points

$$\mathbf{Z} = \begin{pmatrix} z_0 & & 0 \\ & \ddots & \\ 0 & & z_{N-1} \end{pmatrix}$$

through the relation

$$\mathbf{a} = \mathbf{Zh} = \begin{pmatrix} z_0 & & 0 \\ & \ddots & \\ 0 & & z_{N-1} \end{pmatrix} \begin{pmatrix} h_0 \\ \vdots \\ h_{N-1} \end{pmatrix} = \begin{pmatrix} h_0 z_0 \\ \vdots \\ h_{N-1} z_{N-1} \end{pmatrix} = \begin{pmatrix} a_0 \\ \vdots \\ a_{N-1} \end{pmatrix} \in \mathbb{C}^{N\times 1} \tag{6.202}$$

For any diagonal matrix $\mathbf{D}$

$$\mathbf{D} = \operatorname{diag}\left(d_1, \ldots, d_n\right) = \begin{pmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_n \end{pmatrix}$$

it follows that

$$\lambda_i(\mathbf{D}) = d_i, \quad i = 1, \ldots, n \tag{6.203}$$

Our target here is to use the fact of (6.203) for demodulation. Assume that the transmitting signals points remain invariant for n uses of the channel, whose i-th channel gain vector is $\mathbf{h}_i$. So we have

$$\mathbf{A} = \frac{1}{\sqrt{n}}\left(\mathbf{a}_1, \ldots, \mathbf{a}_n\right) = \frac{1}{\sqrt{n}} \begin{pmatrix} z_0 & & 0 \\ & \ddots & \\ 0 & & z_{N-1} \end{pmatrix}_{N\times N} \tag{6.204}$$

$$\begin{pmatrix} h_{00} & h_{01} & \cdots & h_{0,n-1} \\ h_{10} & h_{11} & \cdots & h_{1,n-1} \\ \vdots & \vdots & \vdots & \vdots \\ h_{N-1,0} & h_{n-1,1} & \cdots & h_{N-1,n-1} \end{pmatrix}_{N\times n} = \mathbf{ZH} \in \mathbb{C}^{N\times n}$$

where

$$\mathbf{H} = \frac{1}{\sqrt{n}}\left(\mathbf{h}_1, \ldots, \mathbf{h}_n\right) \in \mathbb{C}^{N\times n}$$

is the channel gain matrix. When $n = N$, we have the eigenvalue decomposition

$$\mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H = \mathbf{U} \begin{pmatrix} \lambda_0(\mathbf{H}) & & 0 \\ & \ddots & \\ 0 & & \lambda_{N-1}(\mathbf{H}) \end{pmatrix} \mathbf{U}^H \tag{6.205}$$

where $\mathbf{U}$ is the matrix consisting of the n eigenvectors and $\lambda_i(\mathbf{H})$ is the i-th (complex) eigenvalue, where $i = 0, \ldots, N-1$. Upon inserting (6.205) into (6.204), we obtain

$$\begin{aligned} \mathbf{A} &= \begin{pmatrix} z_0 & & 0 \\ & \ddots & \\ 0 & & z_{N-1} \end{pmatrix} \mathbf{U} \begin{pmatrix} \lambda_0(\mathbf{H}) & & 0 \\ & \ddots & \\ 0 & & \lambda_{N-1}(\mathbf{H}) \end{pmatrix} \mathbf{U}^H \\ &= \mathbf{U} \begin{pmatrix} z_0 & & 0 \\ & \ddots & \\ 0 & & z_{N-1} \end{pmatrix} \begin{pmatrix} \lambda_0(\mathbf{H}) & & 0 \\ & \ddots & \\ 0 & & \lambda_{N-1}(\mathbf{H}) \end{pmatrix} \mathbf{U}^H \\ &= \mathbf{U} \begin{pmatrix} \lambda_0(\mathbf{H}) z_0 & & 0 \\ & \ddots & \\ 0 & & \lambda_{N-1}(\mathbf{H}) z_{N-1} \end{pmatrix} \mathbf{U}^H, \end{aligned} \tag{6.206}$$

In other words, we have

$$\lambda_i(\mathbf{A}) = \lambda_i(\mathbf{H})\, z_i, \quad i = 0, \ldots, N-1 \tag{6.207}$$

The second line of (6.206) follows from using the special property of the diagonal matrix $\mathbf{Z}$: For any diagonal matrix $\mathbf{D}$ and an arbitrary complex square matrix $\mathbf{C}$, the order of the matrix multiplication can be exchanged

$$\mathbf{DC} = \mathbf{CD}$$

From (6.207) it follows that

$$z_i = \frac{\lambda_i(\mathbf{A})}{\lambda_i(\mathbf{H})}, \quad i = 0, \ldots, N-1 \tag{6.208}$$

Since $\lambda_i(\mathbf{A})$ can be obtained from the outliers of the eigenvalues $\lambda_i\left(\frac{1}{\sqrt{n}}\mathbf{X} + \mathbf{A}\right)$, as pointed out before, we can detect the symbols z_i, $i = 0, \ldots, N-1$, if we know the eigenvalues $\lambda_i(\mathbf{A})$. ∎

Example 6.16.2 (additive Gaussian deformation of the circular law) For each integer $n \geq 1$, let $\mathbf{X} = \left(X_{ij}\right)_{1 \leqslant i,j \leqslant n}$ be the random matrix whose entries are i.i.d. copies of a complex valued random variable ξ with variance σ^2. The circular law theorem asserts that the empirical spectral distribution of $\mathbf{X}$–after centering and rescaling by $\sigma\sqrt{n}$–converges weakly to the uniform distribution on the unit disc of $\mathcal{C}$. Bordenave *et al.* (2013) [334] consider random matrix of the form

$$\mathbf{L} = \mathbf{X} - \mathbf{D} \tag{6.209}$$

where $\mathbf{X}$ is a matrix with i.i.d. entries as above, and $\mathbf{D}$ is the diagonal matrix obtained from the row sums of $\mathbf{X}$: for $i = 1, \ldots, n$,

$$D_{ii} = \sum_{k=1}^{n} X_{ik}$$

If $\mathbf{X}$ is interpreted as the adjacency matrix of a weighted oriented graph, then $\mathbf{L}$ is the associated Laplacian matrix, with zero row sums. ∎

6.17 Random SVD, Single Ring Law, and Outliers

We know that, most times, if one adds a finite rank perturbation to a large random matrix, it barely modifies its spectrum. However, we observe that the extreme eigenvalues may be altered and deviated away from the bulk. This phenomenon has already been well understood in the Hermitian case (see [138] for the references along this line). For a large random Hermitian matrix, if the strength of the added perturbation is above a threshold, the extreme eigenvalues of the perturbed matrix deviate at a macroscopic distance from the bulk (such eigenvalues are usually called outliers) and

2 xd comment: Zhang, *Matrix Theory*, P. 227. We can use something like

$$\lambda_i(\mathbf{AB}) = \lambda_i(\mathbf{A})\,\lambda_i(\mathbf{B})$$

then have Gaussian fluctuations, otherwise they stick to the bulk and fluctuate like those of the nonperturbed matrix. This phenomenon is called the BBP phase transition, named after the authors of [335], who first brought it to light for empirical covariance matrices. Tao [333] studied a non-Hermitian case : he showed in [333] a similar result for large random matrices whose entries are i.i.d. with mean zero and variance one. We have treated the case studied by Tao [333] in Section 6.16. In this section, we study finite rank perturbations for another natural model of non-Hermitian random matrices, which admit a form like

$$\mathbf{X} = \mathbf{U}\begin{pmatrix} s_1 & & 0 \\ & \ddots & \\ 0 & & s_n \end{pmatrix}\mathbf{V} \tag{6.210}$$

where $\mathbf{U}$ and $\mathbf{V}$ are Haar-distributed unitary random matrices and the s_is are positive numbers, which are independent from $\mathbf{U}$ and $\mathbf{V}$ and with additional assumption $(A1)$ the empirical distribution of s_is tends to a probability measure μ_s, which is *compactly supported* on $\mathbb{R}^+$. (6.210) is called random singular value decomposition (SVD). $\mathbf{X}$ is unitary invariant by construction. The complex Ginibre ensemble is a special case of this unitary invariant model.

Example 6.17.1 (Marchenko–Pastur quarter circular law) Let $\mathbf{X}$ be a random matrix with i.i.d. entries. We know that the singular values $s_i(\mathbf{X})$ is compactly supported on $\mathbb{R}^+$. The spectrum measure μ_s for the random matrix $\mathbf{X}$ is the well-known Marchenko–Pastur quarter circular law

$$\mu_s(dx) = \frac{1}{\pi}\sqrt{4 - x^2}\mathbf{1}_{[0,2]}(x)dx$$

where $\mathbf{1}_{[c,d]}(x)$ is the indication function that is one on the interval $[c, d]$ and zero outside the interval. □

Due to Example 6.17.1, the model (6.210) can be seen as a generalization of the i.i.d. matrices case where we assume that they are isotropic, which means that their law does not change by left or right product by any unitary matrix. For example, matrices from the complex Ginibre ensemble (matrices with i.i.d. entries which are complex standard Gaussian) do satisfy the assumption $(A1)$. In [137], Guionnet, Krishnapur and Zeitouni showed that the eigenvalues of $\mathbf{X}$ tend to spread over a single annulus centered in the origin as the dimension n tends to infinity. In [336], Guionnet and Zeitouni proved the convergence in probability of the support of its empirical spectral distribution (ESD), which shows the lack of natural outliers for this kind of matrix. See also (8.7) for the definition of the ESD.

Inspired by Tao [333], Benaych-George and Rochet (2013) [138] proved that, for finite rank perturbation with a bounded operator norm, outliers show up close to the first eigenvalues of the perturbation, which are outside the annulus (as in the i.i.d. matrices case), whereas no outlier appears inside the bulk. Then they showed (and this is the main difficulty of the paper) that the outliers have Gaussian fluctuations in the case of a rank one perturbation.

Let, for each $n \geq 1$, $\mathbf{X}_n$ be a random matrix which admits the decomposition

$$\mathbf{X}_n = \mathbf{U}_n\mathbf{S}_n\mathbf{V}_n$$

with $\mathbf{S}_n = \text{diag}\left(s_1, \ldots, s_n\right)$, where the s_is are positive numbers and where $\mathbf{U}_n$ and $\mathbf{V}_n$are two independent random unitary matrices which are Haar-distributed independently from the matrix $\mathbf{S}_n$. We make the assumptions of the *Single Ring Theorem* [137].

- **Hypothesis 1**: The Empirical Spectral Distribution (ESD) of $\mathbf{S}_n$, $\mu_{\mathbf{S}_n} = \frac{1}{n}\sum_{i=1}^{n}\delta_{s_i}$ converges, in probability, weakly to a deterministic probability measure μ which is compactly supported on $\mathbb{R}^+$.
- **Hypothesis 2**: There exists $M > 0$, such that $\mathbb{P}\left(\left\|\mathbf{S}_n\right\|_{\text{op}} > M\right) \to 0$, where the operator norm is denoted by $\|\cdot\|_{\text{op}}$.
- **Hypothesis 3**: There exist some constants $\kappa, \kappa_1 > 0$ such that

$$\text{Im}(z) > n^{-\kappa} \Rightarrow \left|\text{Im}\left(G_{\mu_{\mathbf{S}_n}}(z)\right)\right| \leqslant \kappa_1$$

where G_μ denotes the Stieltjes transform of μ, that is $G_\mu(z) = \int \frac{1}{z-x}\mu\,(dx)$.

There is anther assumption in the *single ring theorem* [137], but [337] showed that it was unnecessary.

According to [336], we know that the ESD $\mu_{\mathbf{X}_n}$ of $\mathbf{X}_n$ converges, in probability, weakly to a deterministic probability measure whose support on the complex plane is a single ring defined as $\{z \in \mathbb{C}, a \leqslant |z| \leqslant b\}$ .The inner radius a and outer radius b of the ring can be easily calculated:

$$a = \left(\int_0^\infty x^{-2} d\mu_s(x)\right)^{-1/2}, \quad b = \left(\int_0^\infty x^2 d\mu_s(x)\right)^{1/2} \tag{6.211}$$

Example 6.17.2 (singular values have a uniform distribution) Consider the case when singular values have a uniform distribution on the interval $[\alpha, \beta]$:

$$\mu_s(dx) = \frac{1}{\beta - \alpha}\mathbf{1}_{[\alpha,\beta]}dx \tag{6.212}$$

for $\beta > \alpha$. It follows from (6.211) that

$$\begin{aligned} a &= \left(\int_0^\infty x^{-2} d\mu_s(x)\right)^{-1/2} = \left(\frac{1}{\beta-\alpha}\int_\alpha^\beta x^{-2}dx\right)^{-1/2} \\ &= \left(\frac{1}{\beta-\alpha}\left(-x^{-1}\Big|_\alpha^\beta\right)\right)^{-1/2} = \sqrt{\alpha\beta} \end{aligned}$$

and

$$\begin{aligned} b &= \left(\int_0^\infty x^2 d\mu_s(x)\right)^{1/2} = \left(\frac{1}{\beta-\alpha}\int_\alpha^\beta x^2 dx\right)^{1/2} \\ &= \left(\frac{1}{\beta-\alpha}\frac{1}{3}x^3\Big|_\alpha^\beta\right)^{1/2} = \left(\frac{1}{3(\beta-\alpha)}\left(\beta^3 - \alpha^3\right)\right)^{1/2} \\ &= \frac{1}{\sqrt{3}}\sqrt{\alpha^2 + \alpha\beta + \beta^2} \end{aligned}$$

□

According to [336], we know that there is no natural outlier outside the outer circle of the bulk as long as the operator norm $\left\|\mathbf{S}_n\right\|_{\text{op}}$ is bounded, even if $\mathbf{S}_n$ has his own outliers.

Below, to make also sure there is no natural outlier inside the inner circle (when $a > 0$), we may suppose in addition that

$$\sup_{n\geqslant 1} \left\|\mathbf{S}_n^{-1}\right\|_{\text{op}} < \infty$$

Let us now consider a sequence of matrices $\mathbf{A}_n$ (possibly random, but independent of $\mathbf{U}_n$, $\mathbf{S}_n$ and $\mathbf{V}_n$) with rank lower than a fixed integer r such that $\left\|\mathbf{A}_n\right\|_{\text{op}}$ is also bounded. Then, we have:

Theorem 6.17.3 (outliers for finite rank perturbation [138]) Let $\varepsilon > 0$ and suppose that for all sufficiently large n, $\mathbf{A}_n$ does not have any eigenvalues in the band $\{z \in \mathbb{C}, b + \varepsilon \leqslant |z| \leqslant b + 3\varepsilon\}$ and has $k \leq r$ eigenvalues $\lambda_1\left(\mathbf{A}_n\right), \ldots, \lambda_k\left(\mathbf{A}_n\right)$ with modulus higher than $b + 3\varepsilon$. Then, with a probability tending to 1, $\mathbf{X}_n + \mathbf{A}_n$ has exactly k eigenvalues with modulus higher than $b + 2\varepsilon$. Furthermore, after labeling properly

$$\forall i \in \{1, \ldots, k\}, \quad \lambda_i\left(\mathbf{X}_n + \mathbf{A}_n\right) - \lambda_i\left(\mathbf{A}_n\right) \xrightarrow{(\mathbb{P})} 0$$

This theorem is an analogous version of Theorem 1.4 of Tao's paper [333] and so is its proof. However, things are different inside the annulus. Indeed, the following result establishes the lack of small outliers:

Theorem 6.17.4 (no outlier inside the bulk [138]) Suppose that $a > 0$ and $\sup_{n\geqslant 1} \left\|\mathbf{S}_n^{-1}\right\|_{\text{op}} < \infty$. Then for all $\delta \in]0, a[$, with a probability tending to one

$$\mu_{\mathbf{X}_n+\mathbf{A}_n}(\{z \in \mathbb{C}, |z| \leqslant a - \varepsilon\}) = 0$$

where $\mu_{\mathbf{X}_n+\mathbf{A}_n}$ is the Empirical Spectral Distribution of $\mathbf{X}_n + \mathbf{A}_n$.

See Figure 6.32 for an illustration of Theorem 6.17.3 and Theorem 6.17.4. We drew circles around each eigenvalues of $\mathbf{A}_n$ and we observe the lack of outliers inside the annulus.

Let us now consider the fluctuations of the outliers. Here we shall suppose that $\mathbf{A}_n$ has *rank one* and write

$$\mathbf{A}_n = \mathbf{b}_n \mathbf{c}_n^*$$

with $\mathbf{b}_n$ and $\mathbf{c}_n$ some $n \times 1$ complex matrices. We can suppose $\mathbf{c}_n$ to be normalized, so that

$$\left\|\mathbf{A}_n\right\|_{\text{op}} = \sqrt{\mathbf{b}_n \mathbf{b}_n^*}$$

We assume that $\mathbf{A}_n$ has one non zero eigenvalue, which is equal to

$$\theta_n \left(= \mathbf{c}_n^* \mathbf{b}_n\right)$$

whose absolute value $|\theta_n|$ tends to a limit denoted by $|\theta_n|$ such that

$$|\theta| \geqslant b + 4\varepsilon$$

for a certain $\varepsilon > 0$ (b is the radius of the outer circle of the bulk), and we also assume that the largest singular value $\sqrt{\mathbf{b}_n \mathbf{b}_n^*}$ of $\mathbf{A}_n$ converges in the $\mathcal{L}_2$ sense to a limit denoted by L when n goes to $+\infty$.

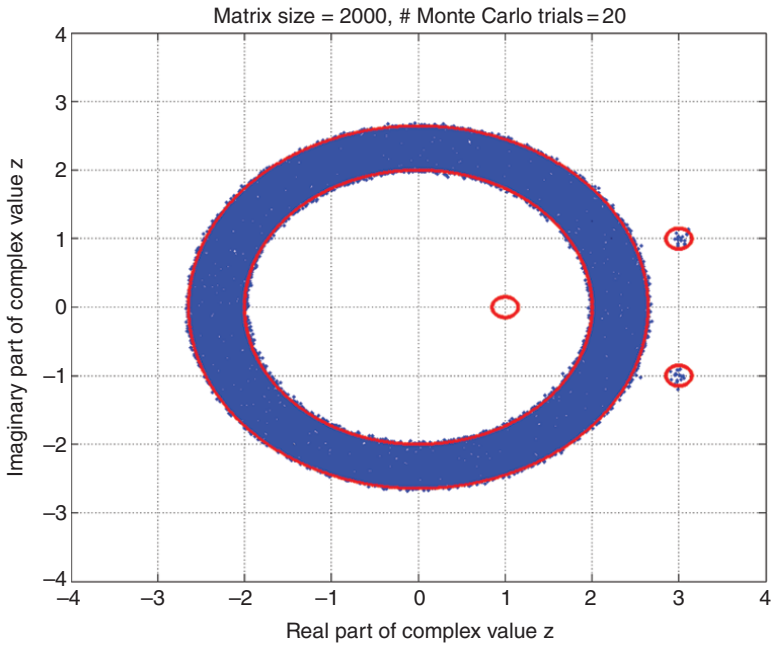


Figure 6.32 Eigenvalues of $\mathbf{X}_n + \mathbf{A}_n$ for $n = 2000$ where $\mu_s(dx) = \frac{1}{3}\mathbf{1}_{[1,4]}(x)dx$ and $\mathbf{A}_n = \text{diag}(1, 3+i, 3-i, 0, \ldots, 0)$. The small circles are centered at $1, 3+i, 3-i$, respectively, and each has a radius of $\sim \frac{1}{n^{1/4}}$. Twenty Monte Carlo trials are performed.

It turns out that the fluctuations of the outlier eigenvalue of $\mathbf{X}_n + \mathbf{A}_n$ are Gaussian with a variance that can be explicitly expressed out of L, θ and b (see (6.213)). More precisely, we have:

Theorem 6.17.5 (Gaussian fluctuations away from the bulk [138]) Let $\tilde{\lambda}_n$ denote the largest eigenvalue of $\mathbf{X}_n + \mathbf{A}_n$ in absolute value. Then as n tends to infinity

$$\sqrt{n}\left(\tilde{\lambda}_n - \theta_n\right) \xrightarrow{(d)} \mathcal{N}_{\mathbb{C}}\left(0, \frac{b^2 L^2}{|\theta|^2 - b^2}\right) \tag{6.213}$$

where $\mathcal{N}_{\mathbb{C}}\left(0, \sigma^2\right)$ denotes the complex Gaussian law with covariance matrix $\sigma^2 \mathbf{I}_2$.

We notice that the closer from the bulk $|\theta|$ is, the higher the variance is. This can be viewed as a new expression of the general tendency of eigenvalues of random matrices to repel each other. However, the limit law is isotropic in $\mathbb{C}$, whereas it could have been plausible that this tendency would have led to get a nonsymmetric limit distribution here, due to a tendency of $\tilde{\lambda}_n$ to get as far as possible from the bulk.

Due to [336], we know that Theorem 6.17.5 applies, for example, to the model of random complex matrices $\mathbf{X}_n$ distributed according to the law

$$\frac{1}{Z_n} \exp\left(-n \operatorname{Tr} V\left(\mathbf{A}\mathbf{A}^*\right)\right) d\mathbf{A}$$

where $d\mathbf{A}$ is the Lebesgue measure of the $n \times n$ complex matrices set, $V(x)$ is a polynomial with positive leading coefficient and Z_n is a normalization constant. It is quite a natural unitarily invariant model. One can notice that $V(x) = \frac{1}{2\sigma^2}x$ gives the Ginibre matrices [111, 328, 329].

Example 6.17.6 (massive MIMO in the presence of interference and noise) For the sake of conciseness, let the channel bandwidth be smaller than the coherence bandwidth. Channels whose physical bandwidth is wider than the coherence bandwidth can be decomposed into equivalent parallel narrowband channels by means of orthogonal frequency division multiplexing or related techniques.

Let the frequency-flat, block-fading, narrowband channel from m transmit antennas to n receive antennas be described by the matrix equation

$$\mathbf{Y} = \mathbf{HT} + \mathbf{Z} \tag{6.214}$$

where $\mathbf{T} \in \mathbb{C}^{m\times T}$ is the transmitted data (eventually multiplexed with pilot symbols), T is the coherence time in multiples of the symbol interval, $\mathbf{H} \in \mathbb{C}^{n\times m}$ is the channel matrix of unknown propagation coefficients, $\mathbf{Y} \in \mathbb{C}^{n\times T}$, is the received signal, and $\mathbf{Z} \in \mathbb{C}^{m\times T}$ is the total impairment. Furthermore, we assume that channel, data, and impairment have zero mean, i.e. $\mathbb{E}\mathbf{X} = \mathbb{E}\mathbf{H} = \mathbb{E}\mathbf{Z} = \mathbf{0}$. The impairment includes both thermal noise and interference from other cells and is, in general, neither white nor Gaussian.

We decompose the impairment process

$$\mathbf{Z} = \mathbf{W} + \mathbf{H}_I\mathbf{X}_I \tag{6.215}$$

into white noise $\mathbf{W}$ and interference from L neighboring cells where interfering data $\mathbf{X}_I \in \mathbb{C}^{Lm\times n}$ is transmitted in neighboring cells and received in the cell of interest through the channel $\mathbf{H}_I \in \mathbb{C}^{n\times Lm}$. It follows from (6.214) and (6.215) that

$$\mathbf{Y} = \mathbf{HT} + \mathbf{H}_I\mathbf{X}_I + \mathbf{W} \tag{6.216}$$

The interference and the noise can be modeled using (6.210) when

$$\mathbf{Z} = \mathbf{X} = \mathbf{U}^H \begin{pmatrix} s_1 & & 0 \\ & \ddots & \\ 0 & & s_n \end{pmatrix} \mathbf{V}$$

Recognizing $\mathbf{A} = \mathbf{HT}$, we reach the standard model $\mathbf{X} + A$, which is considered above. $\mathbf{A}$ is the low rank matrix whose rank r is $r = \min\{m, n, T\}$. For example if we set the parameters $m = 3, T = 1000, L = 2, n = 200$, we have the rank $r = 3$. Consider the special case of the single transmit antenna $m = 1$. We have the rank one matrix perturbation, $\mathbf{h}\mathbf{t}^T$, of the random matrix $\mathbf{Z}$

$$\mathbf{Y} = \mathbf{h}\mathbf{t}^T + \mathbf{Z}$$

where $\mathbf{h} = \mathbf{H} \in \mathbb{C}^{n\times 1}, \mathbf{t} = \mathbf{T}^T \in \mathbb{C}^{T\times 1}$. Only one outlier will occur in the spectrum of $\mathbf{Y} = \mathbf{h}\mathbf{t}^T + \mathbf{Z}$ in the complex plane. For m transimit antennas we have

$$\mathbf{Y} = \mathbf{h}_1\mathbf{t}_1{}^T + \mathbf{h}_2\mathbf{t}_2{}^T + \cdots + \mathbf{h}_m\mathbf{t}_m{}^T + \mathbf{Z}$$

where $\mathbf{h}_i = \mathbf{H}(:, i) \in \mathbb{C}^{n\times 1}, \mathbf{t}_i = \mathbf{T}^T(:, i) \in \mathbb{C}^{T\times 1}, i = 1, \ldots, m,$ if $m \le n,$ and $m \le T.$ There are $r(= m)$ outliers for this case. □

Code 3: Outliers in the Single Ring Theorem

```
%*************************************************************
%
% Outliers in the single ring theorem
%
%          FLORENT BENAYCH-GEORGES AND JEAN ROCHET
%
%   arXiv: 1308.3064v1 [math.PR] 14 Aug 2013
%*************************************************************
clear all;
n=200; % matrix of n x n
N_Try=20 % number of Monte Carlo trials
Axis_Length=4; % window of visualization
IFIG=0;

A=zeros(n,n);
alpha=1; beta=4;

A(1,1)=1; A(2,2)=3+i;A(3,3)=3-i;

for i=1:N_Try
X=zeros(n,n);
X=(randn(n,n)+j*randn(n,n))/sqrt(2);
% i.i.d. random matrix with i.i.d.
 %(Gaussian) complex entries
[U1,S1,V1] = svd(X);
X=(randn(n,n)+j*randn(n,n))/sqrt(2);
% i.i.d. random matrix with i.i.d.
 %(Gaussian) complex entries
[U2,S2,V2] = svd(X);
c=alpha;d=beta; s = c+ (d-c).*rand(n,1);
% uniform distribution
 % on the interval [c, d]
S=diag(s); % singular eigenvalues have uniform distrition
on the interval [c,d]
X=U1*S*V2; % random matrix with prescribed singular values
lamda=eig(X+A); % eigenvalues are complex numbers
SNR_dB=10*log10(trace(A*A')/trace(X*X'/n))
%*********** Figures *********
IFIG=1;
IFIG=IFIG+1;figure(IFIG);
t=0:2*pi/1000:2*pi;x=sin(t);y=cos(t); % unit circle
r=n^(-1/4); % radius of circle
a=sqrt(alpha*beta); b=sqrt(alpha^2+beta^2+alpha*beta)/sqrt(3);
plot(real(lamda),imag(lamda),'.', a*x,a*y,'r.',b*x,b*y,'r.',...
3+r*x,1+r*y,'r.',3+r*x,-1+r*y,...
```

```
 'r.',1+r*x,0+r*y,'r.');
hold on;
axis([Axis_Length*(-1) Axis_Length Axis_Length*(-1)
Axis_Length]) grid on;
xlabel('Real Part of Complex Value z')
ylabel('Imaginary Part of Complex Value z')
title(['Matrix size = ',num2str(n), ',
# Monte Carlo Trials=',num2str(i)])
end % N_Try
hold off
```

6.17.1 Outliers for Finite Rank Perturbation: Proof of Theorem 6.17.3

Now we give outline the proof Theorem 6.17.3, with the emphasis of what tools are needed. We adapt [333, Theorem 1.4] to prove Theorem 6.17.3. It starts with this calculation

$$\begin{aligned}\det(z\mathbf{I}-(\mathbf{X}+\mathbf{A})) &= \det(z\mathbf{I}-\mathbf{X})\det\left(\mathbf{I}-(z\mathbf{I}-\mathbf{X})^{-1}\mathbf{A}\right)\\ &= \det(z\mathbf{I}-\mathbf{X})\det\left(\mathbf{I}-(z\mathbf{I}-\mathbf{X})^{-1}\mathbf{BC}\right)\\ &= \det(z\mathbf{I}-\mathbf{X})\det\left(\mathbf{I}-\mathbf{C}(z\mathbf{I}-\mathbf{X})^{-1}\mathbf{B}\right)\end{aligned} \tag{6.217}$$

where $\mathbf{A}=\mathbf{BC}, \mathbf{B}\in\mathbb{C}^{n\times r}, \mathbf{C}\in\mathbb{C}^{r\times n}$. For the last step, we used the fact that for all $\mathbf{M}\in\mathbb{C}^{p\times q}, \mathbf{N}\in\mathbb{C}^{q\times p}$

$$\det\left(\mathbf{I}_p+\mathbf{MN}\right)=\det\left(\mathbf{I}_q+\mathbf{NM}\right)$$

For any matrix $\mathbf{Q}\in\mathbb{C}^{n\times n}$, the n eigenvalues z satisfy $\det\left(z\mathbf{I}_n+\mathbf{Q}\right)=0$.

According to (6.217), therefore, the eigenvalues z of $\mathbf{X}+\mathbf{A}$ are not eigenvalues of $\mathbf{X}$, and they instead satisfy

$$\det\left(\mathbf{I}-\mathbf{C}(z\mathbf{I}-\mathbf{X})^{-1}\mathbf{B}\right)=0.$$

Using (6.217), as previously done by Tao in [333], we introduce the meromorphic functions (implicitly depending on n)

$$f(z) := \det\left(\mathbf{I}-\mathbf{C}(z\mathbf{I}-\mathbf{X})^{-1}\mathbf{B}\right) \tag{6.218}$$

$$g(z) := \det\left(\mathbf{I}-\mathbf{C}(z\mathbf{I})^{-1}\mathbf{B}\right) \tag{6.219}$$

Lemma 6.17.7 As n goes to infinity, we have

$$\sup_{|z|\geq b+2\varepsilon}|f(z)-g(z)| \xrightarrow{\mathbb{P}} 0$$

See the proof of this lemma below. Now we are in a position to explain how this lemma allows one to conclude the proof of Theorem 6.17.3. The poles of $f(z)$ and $g(z)$ are, respectively, eigenvalues of the $\mathbf{A}$ and of the null matrix $\mathbf{I}$, hence for n large enough, they have no pole in the region $\{z\in\mathbb{C} : |z|>b+2\varepsilon\}$, whereas their zeros in this region are precisely the eigenvalues of $\mathbf{X}+\mathbf{A}$ and $\mathbf{A}$ that are in this region. Thus by Rouché's Theorem, with probability tending to 1, for n large enough, $\mathbf{X}+\mathbf{A}$ and $\mathbf{A}$ have the same number j of eigenvalues in this region. Indeed, $|g(z)|$ admits the following lower bound

on the circle with radius $b+\varepsilon$: as we assumed that any eigenvalue of $\mathbf{A}$ is at least at distance ε from $\{z \in \mathbb{C} : |z| = b+2\varepsilon\}$, we have

$$\inf_{|z|=b+2\varepsilon} |g(z)| = \inf_{|z|=b+2\varepsilon} \frac{\prod_{i=1}^{n} |z-\lambda_i(\mathbf{A})|}{|z|^2} \geqslant \left(\frac{\varepsilon}{b+2\varepsilon}\right)^r$$

Also, using Lemma 6.17.7, we conclude that, after a proper labeling

$$\forall i \in \{1,\dots,j\}, \quad \lambda_i(\mathbf{X}+\mathbf{A}) \xrightarrow{\mathbb{P}} \lambda_i(\mathbf{A})$$

Indeed, for each fixed $i \in \{1,\dots,j\}$

$$\prod_{\ell=1}^{n} \left|1 - \frac{\lambda_\ell(\mathbf{A})}{\lambda_i(\mathbf{X}+\mathbf{A})}\right| = \left|g\left(\lambda_i(\mathbf{X}+\mathbf{A})\right)\right| = \left|f\left(\lambda_i(\mathbf{X}+\mathbf{A})\right) - g\left(\lambda_i(\mathbf{X}+\mathbf{A})\right)\right| \leqslant \sup_{|z|\geqslant b+2\varepsilon} |f(z)-g(z)| \xrightarrow{\mathbb{P}} 0$$

Let us now explain how to prove Lemma 6.17.7. we can notice at first that it suffices to prove that

$$\sup_{|z|\geqslant b+2\varepsilon} |f(z)-g(z)| = \sup_{|z|\geqslant b+2\varepsilon} \left|\det\left(\mathbf{I}-\mathbf{C}(z\mathbf{I}-\mathbf{X})^{-1}\mathbf{B}\right) - \det\left(\mathbf{I}-\mathbf{C}(z\mathbf{I})^{-1}\mathbf{B}\right)\right| \xrightarrow{\mathbb{P}} 0$$

simply because the function $\det : \mathbb{C}^{r\times r} \to \mathbb{C}$ is Lipschitz over every bounded set of complex matrices $\mathbb{C}^{r\times r}$. Then, the proof of Lemma 6.17.7 is based on both following lemmas (whose proofs are found in the original source [138]). Let $\|\cdot\|_{\text{op}}$ be the operator norm of the matrix.

Lemma 6.17.8 There exists a constant $C_1 > 0$ such that the event

$$\mathcal{E}_n : \left\{\forall k \geqslant 1, \quad \left\|\mathbf{X}^k\right\|_{\text{op}} \leqslant C_1 \cdot (b+\varepsilon)^k\right\}$$

has probability tending to 1, as n tends to infinity.

Lemma 6.17.9 For all $k \geq 0$, as n tends to infinity, we have

$$\left\|\mathbf{C}\mathbf{X}^k\mathbf{B}\right\|_{\text{op}} \xrightarrow{\mathbb{P}} 0$$

On the event $\mathcal{E}_n$ defined at Lemma 6.17.8 above, we write, for $|z| \geqslant b+2\varepsilon$

$$\mathbf{C}(z\mathbf{I}-\mathbf{X})^{-1}\mathbf{B} - \mathbf{C}(z\mathbf{I})^{-1}\mathbf{B} = \mathbf{C}\sum_{k=1}^{+\infty} \frac{\mathbf{X}^k}{z^{k+1}}\mathbf{B}$$

and it suffices to write that for any $\delta > 0$

$$\mathbb{P}\left(\sup_{|z|\geqslant b+2\varepsilon} \left\|\mathbf{C}(z\mathbf{I}-\mathbf{X})^{-1}\mathbf{B} - \mathbf{C}(z\mathbf{I})^{-1}\mathbf{B}\right\|_{\text{op}} > \delta\right) \leqslant \mathbb{P}\left(\mathcal{E}_n^c\right) + \mathbb{P}\left(\sum_{k=1}^{k_0} \frac{\left\|\mathbf{C}\mathbf{X}^k\mathbf{B}\right\|_{\text{op}}}{(b+2\varepsilon)^{k+1}} > \frac{\delta}{2}\right) + \mathbb{P}\left(\mathcal{E}_n \text{ and } \left\|\mathbf{C}\sum_{k=k_0+1}^{+\infty} \frac{\mathbf{X}^k}{z^{k+1}}\mathbf{B}\right\|_{\text{op}} > \frac{\delta}{2}\right)$$

Due to Lemma 6.17.8, we find a large enough k_0 so that the last event has a vanishing probability. Then, Lemma 6.17.9, the probability of the last-but-one event goes to zero as n tends to infinity.

6.17.2 Eigenvalues Inside the Inner Circle: Proof of Theorem 6.17.4

Our goal here is to show that for all $\delta \in]0, a[$, with probability tending to one, the function $f(z)$ defined at (6.218) has no zero in the region $\{z \in \mathbb{C} : |z| < a - \delta\}$. Recall that

$$f(z) := \det\left(\mathbf{I} - \mathbf{C}(z\mathbf{I} - \mathbf{X})^{-1}\mathbf{B}\right)$$

so that a simple sufficient condition would be $\left\|\mathbf{C}(z\mathbf{I} - \mathbf{X})^{-1}\mathbf{B}\right\|_{\text{op}} < 1$ for all $|z| < a - \delta$. Thus, it suffices to prove that with probability tending to one as n tends to infinity

$$\sup_{|z|<a-\delta} \left\|\mathbf{C}(z\mathbf{I} - \mathbf{X})^{-1}\mathbf{B}\right\|_{\text{op}} < 1$$

The method is the same as in Section 6.17.1. Let us write, for all $|z| < a - \delta$

$$\begin{aligned}\mathbf{C}(z\mathbf{I} - \mathbf{X})^{-1}\mathbf{B} &= \mathbf{C}\mathbf{X}^{-1}\left(z\mathbf{I} - \mathbf{X}^{-1}\right)^{-1}\mathbf{B} \\ &= \mathbf{C}\sum_{k=1}^{+\infty} z^{k-1}\mathbf{X}^{-k}\mathbf{B}\end{aligned} \tag{6.220}$$

The idea is to see $\mathbf{X}^{-1}$ as an isotropic random matrix such as $\mathbf{X}$, since $\mathbf{X}^{-1} = \mathbf{V}^H \operatorname{diag}\left(\frac{1}{s_1}, \ldots, \frac{1}{s_n}\right)\mathbf{U}^H$, and satisfies the same kind of hypothesis.

According to [336], we know that there is no natural outlier outside the outer circle of the bulk as long as $\|\mathbf{S}_n\|_{\text{op}}$ is bounded, even if $\mathbf{S}_n$ has his own outliers. In Theorem 6.17.4, to make also sure there is no natural outlier inside the inner circle (when $a > 0$), we may suppose in addition that $\sup_{n\geqslant 1} \|\mathbf{S}_n^{-1}\|_{\text{op}} < \infty$.

Indeed, Hypotheses 1 and 2 are automatically satisfied because $a > 0$, and the following lemma assures us that Hypothesis 3 is also satisfied.

Lemma 6.17.10 There exist some constants $\tilde{\kappa}, \tilde{\kappa}_1 > 0$, such that

$$\operatorname{Im}(z) > \frac{1}{n^{\tilde{\kappa}}} \Rightarrow \left|\operatorname{Im}\left(G_{\mathbf{S}^{-1}}(z)\right)\right| \leqslant \tilde{\kappa}_1$$

The proof of Lemma 6.17.10 is given in [138]. Thus, according to [336], the support of $\mu_{\mathbf{X}_n^{-1}}(\cdot)$ converges in probability to the annulus

$$\left\{z \in \mathbb{C} : b^{-1} \leqslant |z| < a^{-1}\right\}$$

and so, according to what we have done previously

$$\sup_{|\xi|\geqslant a^{-1}+\varepsilon} \mathbf{C}\sum_{k=1}^{+\infty} \frac{\mathbf{X}^{-k}}{\xi^{k+1}}\mathbf{B} \xrightarrow{\mathbb{P}} 0$$

Therefore

$$\mathbb{P}\left(\sup_{|z|<a-\delta} \left\|\mathbf{C}(z\mathbf{I} - \mathbf{X})^{-1}\mathbf{B}\right\|_{\text{op}} < 1\right) \geqslant 1 - \mathbb{P}\left(\sup_{|\xi|>a^{-1}+\varepsilon} \left\|\mathbf{C}\sum_{k=1}^{+\infty} \frac{\mathbf{X}^{-k}}{\xi^{k+1}}\mathbf{B}\right\|_{\text{op}} < 1\right) \to 1$$

with a proper choice for ε.

6.18 The Elliptic Law and Outliers

Let us consider an array of random variables $X_{ij}, 1 \leqslant i,j < \infty$, such that the pairs $\left(X_{ij}, X_{ji}\right), 1 \leqslant i \leqslant j < \infty$, are independent random vectors with zero mean $\mathbb{E}X_{ij} = \mathbb{E}X_{ji} = 0$, variance one $\mathbb{E}X_{ij}^2 = \mathbb{E}X_{ji}^2 = 1$, and correlation coefficient $\mathbb{E}X_{ij}X_{ji} = \rho, |\rho| \leqslant 1$. We also assume that $X_{ii}, i \leqslant i < \infty$, are independent random variables, independent of the pairs $\left(X_{ij}, X_{ji}\right), 1 \leqslant i \leqslant j < \infty$, and $\mathbb{E}X_{ii} = 0, \mathbb{E}X_{ii}^2 < \infty$. We consider the random matrix

$$\mathbf{X}_n = \left\{X_{ij}\right\}_{i,j=1}^n$$

Define the empirical spectral measure μ_n of $n^{-1/2}\mathbf{X}_n$ by (3.6).

Theorem 6.18.1 (elliptic law) Let $\mathbf{X}_n$ be given above. Then $\mu_n \to \mu$ in probability, and μ has the density $g(x,y)$:

$$g\left(x,y\right) = \begin{cases} \frac{1}{\pi(1-\rho^2)}, & \left(x,y\right) \in \left\{u,v \in \mathbb{R} : \quad \frac{u^2}{(1+\rho)^2} + \frac{v^2}{(1+\rho)^2} \leqslant 1\right\} \\ 0, & \text{otherwise} \end{cases}$$

Example 6.18.2 (MATLAB implementations) Consider two standard Gaussian random variables, Y and Z. We form the random matrix $\mathbf{X}_n$ using

$$X_{ij} = Y; \quad X_{ji} = \rho Y + \sqrt{1-\rho^2}Z, \quad 1 \leqslant i < j < n \tag{6.221}$$

Another method in MATLAB is two generate two random vectors with correlation coefficient ρ with covariance matrix

$$\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

The trick is to use the Cholesky factorization function `chol.m`. If $\mathbf{A}$ is positive definite, `R = chol(A)` produces an upper triangular $\mathbf{R}$ such that

$$\mathbf{R}^T\mathbf{R} = \mathbf{A}$$

```
R = chol(Sigma);
for i=1:n % since we deal with i.i.d. random variables,
we use the loop.
 z = repmat(mu,n,1) + bernoulli(0.5,n,2)*R;
 % z is a n x 2 matrix
 for j=1:n;
X(i,j)=z(j,1); % the first random vector
X(j,i)=z(j,2); % the second random vector that is correlated
               % with the first random vector
end
end
```

□

Suppose $\mathbf{A}$ is an $n \times n$ matrix

Denote by $\lambda_1, \ldots, \lambda_n$ the eigenvalues of the matrix $\mathbf{X}_n$ and define its spectral distribution function $F_{\mathbf{A}_n}(x,y)$ by (3.6).

If $\rho = 1$, we have the ensemble of symmetric random matrices. If X_{ij} are i.i.d., then $\rho = 0$, and we get the ensemble of matrices with i.i.d. elements.

Define the density of uniformly distributed random variable on the ellipse

$$g(x,y) = \begin{cases} \frac{1}{\pi(1-\rho^2)}, & (x,y) \in \left\{ u,v \in \mathbb{R} : \quad \frac{u^2}{(1+\rho)^2} + \frac{v^2}{(1+\rho)^2} \leqslant 1 \right\} \\ 0, & \text{otherwise,} \end{cases}$$

and the corresponding distribution function

$$G(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u,v)\, dudv$$

If all X_{ij} have finite fourth moment and densities then it was proved by Girko that $F_{\mathbf{X}_n}$ converges to G. He called this result "elliptic law." But similarly to the case of the Circular law Girko's proof is considered questionable in the literature. Later the elliptic law was proved for matrices with Gaussian entries [338]. In this case one can write an explicit formula for the density of eigenvalues of the matrix $n^{-1/2}\mathbf{X}_n$. In [339, 340], Naumov proved that the elliptic law under the assumption that all elements have a finite fourth moment only. More recently, Nguyen and O'Rourke [341] proved proved the elliptic law in the general case assuming finite second moment only. The line of work is relevant to the circular law (see tutorial [329]).

Figure 6.33 is illustrated for a Gaussian random variable while Figure 6.34 for Bernoulli random variable. Both are for $\rho = 0.5$. Figure 6.35 and Figure 6.36 are for $\rho = -0.5$.

Example 6.18.3 (Gaussian case) Let the elements of the matrix $\mathbf{X}$ have Gaussian distribution with zero mean and correlations

$$\mathbb{E}X_{ij}^2 = 1, \quad \text{and} \quad \mathbb{E}X_{ij}X_{ji} = \rho, \quad i \neq j, \quad |\rho| \leqslant 1$$

The ensemble of such matrices can be specified by the probability measure

$$\mathbb{P}(dX) \sim \exp\left[-\frac{n}{2(1-\rho^2)} \operatorname{Tr}\left(XX^T - \rho X^2 \right) \right]$$

It follows from Theorem 6.18.1 that $\mu_n \xrightarrow{\text{weak}} \mu$. This result can be generalized to the ensemble of Gaussian complex asymmetric matrices. In this case, the invariant measure is

$$\mathbb{P}(dX) \sim \exp\left[-\frac{n}{1-|\rho|^2} \operatorname{Tr}\left(XX^T - 2\operatorname{Re}\rho X^2 \right) \right] \tag{6.222}$$

and

$$\mathbb{E}X_{ij}^2 = 1, \quad \text{and} \quad \mathbb{E}X_{ij}X_{ji} = |\rho|\, e^{j2\theta}, \quad i \neq j, \quad |\rho| \leqslant 1$$

Then the limit measure has a uniform density inside an ellipse which is centered at zero and has semiaxes $1 + |\rho|$ in the direction θ and $1 - |\rho|$ in the direction $\theta + \pi/2$. For illustration, see Figure 6.37 (Gaussian case) and Figure 6.38 (Bernoulli case).

It seems that (6.222) is also valid for the Bernoulli case, according to Figure 6.38. □

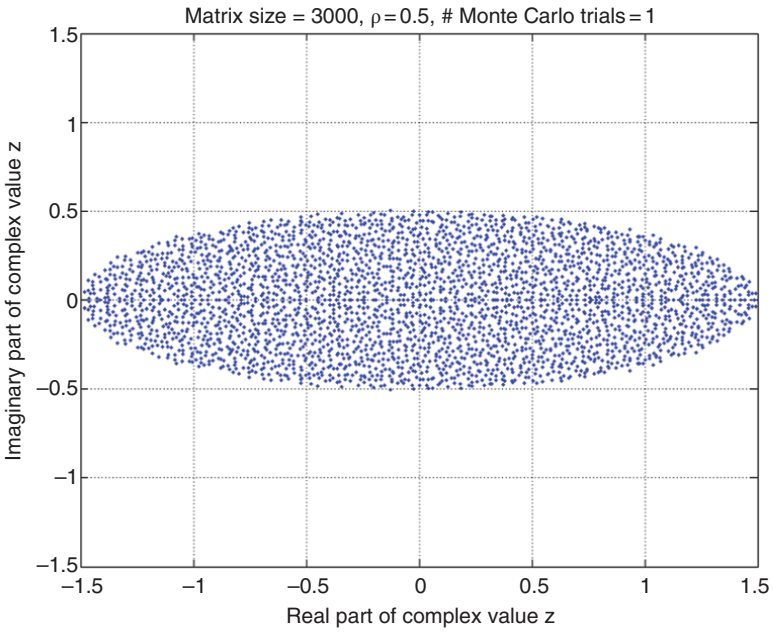


Figure 6.33 Eigenvalues of the matrix $n^{-1/2}\mathbf{X}_n$ for $\rho = 0.5$ and $n = 3000$. Each entry is an i.i.d. Gaussian random variable.

In analogy with the outliers in the circular law in Section 6.16, here we conjecture that outliers also will occur in $n^{-1/2}\mathbf{X}_n + \mathbf{A}_n$ where $\mathbf{A}_n$ is an $n \times n$ matrix of low rank. Formal proof of this result is beyond the reach of the author. The outliers of $n^{-1/2}\mathbf{X}_n + \mathbf{A}_n$ is illustrated in Figure 6.39.

Code 4: Outliers in the Elliptic Law

```
%************************************************************
%
% ELLIPTIC LAW FOR REAL RANDOM MATRICES
%
%       ALEXEY NAUMOV
%
% arXiv: 1201.1639v2 [math.PR] 13 Feb 2012
%************************************************************
clear all;
n=3000; % matrix of n x n
N_Try=1 % number of Monte Carlo trials
Axis_Length=1.5; % window of visualization
IFIG=0;
rho=0.5;
A=zeros(n,n);X=zeros(n,n);
 A(1,1)=1; A(2,2)=2.5+i;A(3,3)=2.5-i;
```

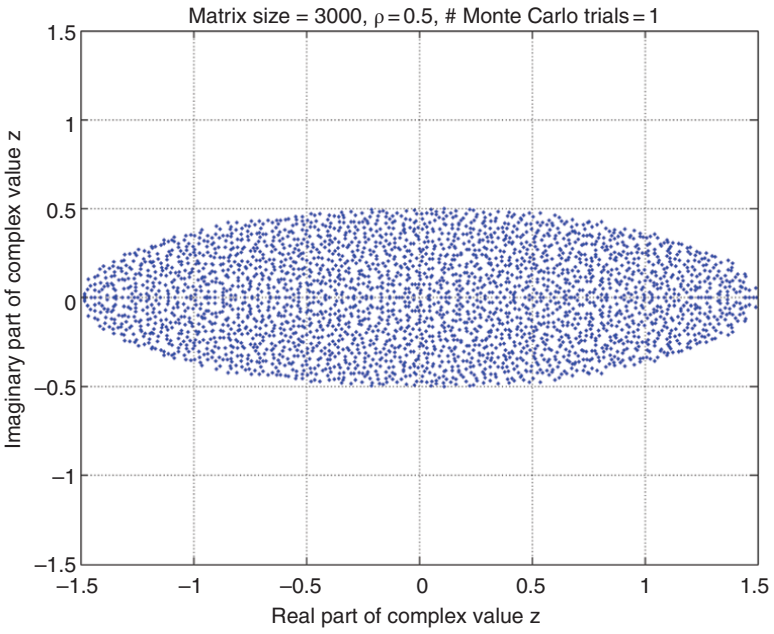


Figure 6.34 Eigenvalues of the matrix $n^{-1/2}\mathbf{X}_n$ for $\rho = 0.5$ and $n = 3,000$. Each entry is an i.i.d. Bernoulli random variable, taking the values +1 and −1 each with probability 1/2.

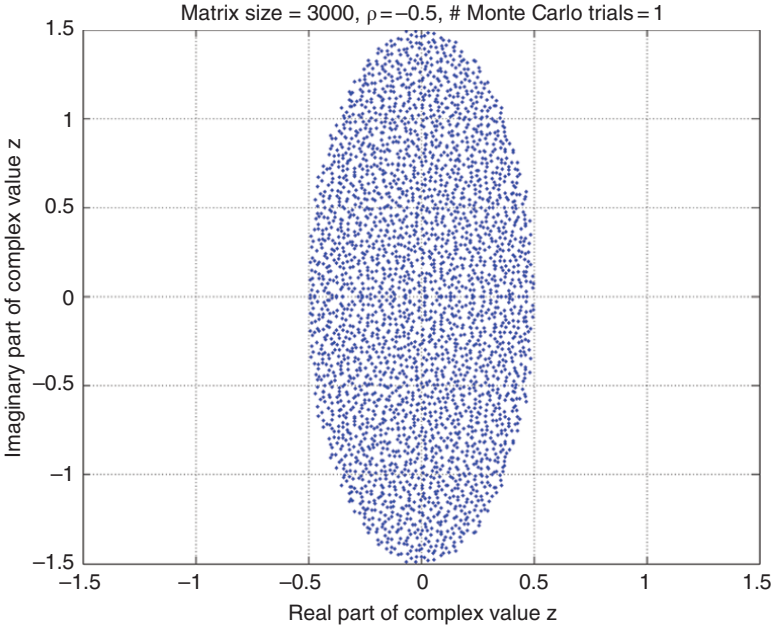


Figure 6.35 Eigenvalues of the matrix $n^{-1/2}\mathbf{X}_n$ for $\rho = -0.5$ and $n = 3000$. Each entry is an i.i.d. Gaussian random variable.

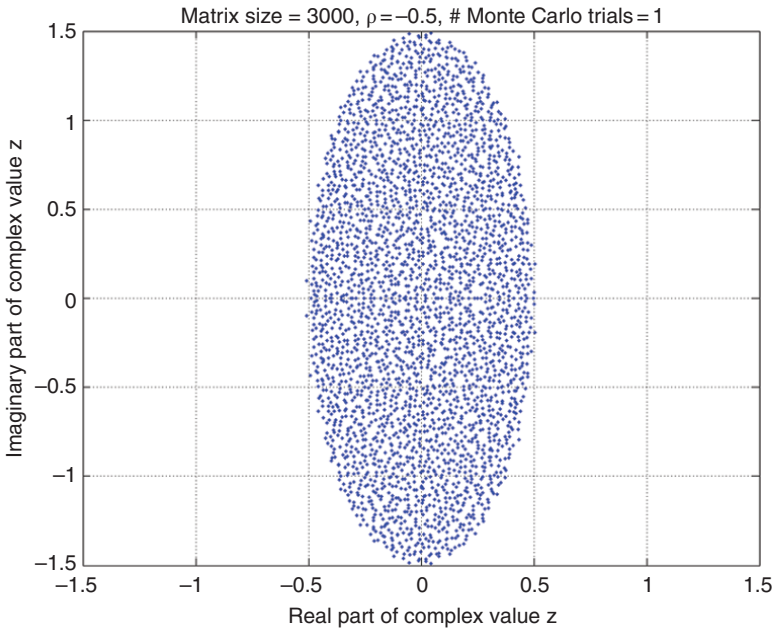


Figure 6.36 Eigenvalues of the matrix $n^{-1/2}\mathbf{X}_n$ for $\rho = -0.5$ and $n = 3000$. Each entry is an i.i.d. Bernoulli random variable, taking the values +1 and −1 each with probability 1/2.

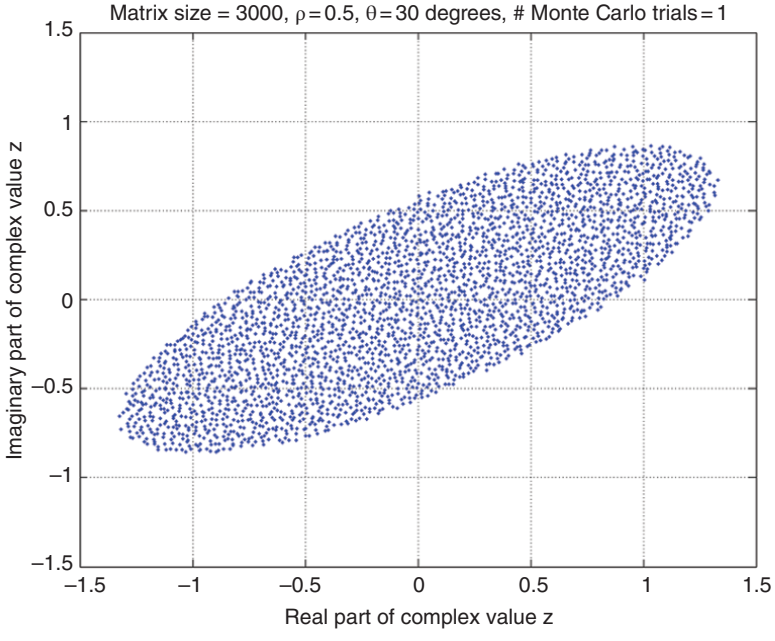


Figure 6.37 Eigenvalues of the matrix $n^{-1/2}\mathbf{X}_n$ for $\rho = -0.5$, $\theta = 30$ and $n = 3000$. Each entry is an i.i.d. Gaussian random variable.

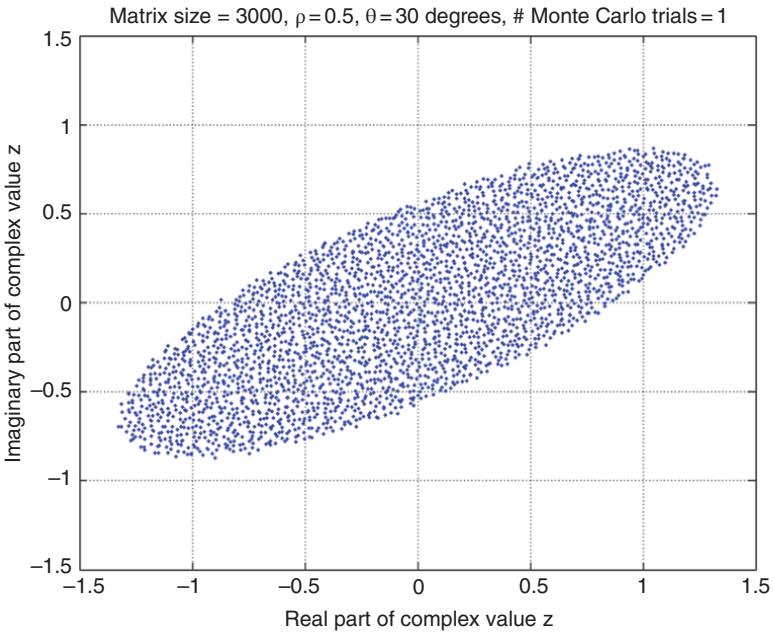


Figure 6.38 Eigenvalues of the matrix $n^{-1/2}\mathbf{X}_n$ for $\rho = -0.5$, $\theta = 30$ and $n = 3,000$. Each entry is an i.i.d. Bernoulli random variable, taking the values +1 and −1 each with probability 1/2.

```
for i_Try=1:N_Try
D=randn(n,1);
%Method 1 for generating two correlated random variables
mu = [0 0];
Sigma = [1 rho; rho 1]; R = chol(Sigma);

%Method 2 for generating two correlated random variables
%for i=1:n
 %for j=1:n
%X(i,j)=randn(1,1);
%X(j,i)=rho *X(i,j) + sqrt(1-rho^2)*randn(1,1);
%end
%end
for i=1:n
 % Gaussian random variable
z = repmat(mu,n,1) + randn(n,2)*R;
 % Bernoulli random variable
%z = repmat(mu,n,1) + bernoulli(0.5,n,2)*R;
 j=1:n;
X(i,j)=z(j,1);
X(j,i)=z(j,2);
end
for i=1:n
```

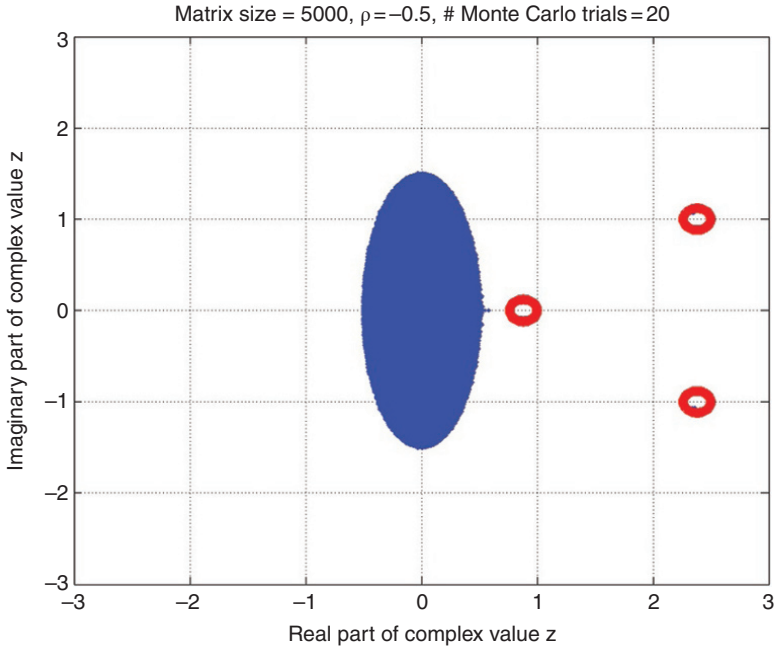


Figure 6.39 Plotted above is the distribution of the eigenvalues of $n^{-1/2}\mathbf{X}_n + \mathbf{A}_n$ where $\mathbf{X}_n$ is an $n \times n$ random matrix with $n = 3000$ and $\rho = 0.5$. Each entry of $\mathbf{X}_n$ is an i.i.d. Gaussian random variable. $\mathbf{A}_n = \text{diag}\,(1, 2.5 + i, 2.5 - i, 0, \ldots, 0)$. The three circles with radius of $1/n^{1/4}$ are located at $1, 2.5 + i, 2.5 - i$. Twenty Monte Carlo trials are performed.

```
 X(i,i)=D(i); % diagonal elments are zero mean with finite
    variance
end

lamda=eig(X/sqrt(n)+A); % eigenvalues are complex numbers
SNR_dB=10*log10(trace(A*A')/trace(X*X'/n))
%*********** Figures *********
IFIG=0;
IFIG=IFIG+1;figure(IFIG);
t=0:2*pi/1000:2*pi;x=sin(t);y=cos(t); % unit circle
r=n^(-1/4); % radius of circle

plot(real(lamda),imag(lamda),'.',1-r+r*x,r*y,'r*',
2.5+-r+r*x,1+r*y,'r*',...
2.5-r+r*x,-1+r*y,'r*')
hold on;
axis([Axis_Length*(-1) Axis_Length Axis_Length*(-1)
    Axis_Length])
grid on;
xlabel('Real Part of Complex Value z')
```

```
ylabel('Imaginary Part of Complex Value z')
title(['Matrix size = ',num2str(n),', \rho= ',num2str(rho),
', # Monte Carlo Trials=',num2str(i_Try)])
end % N_Try
hold off


function B=bernoulli(p,m,n);
% BERNOULLI.M
% This function generates n independent draws of a Bernoulli
% random variable with probability of success p.
% first, draw n uniform random variables

M = m;
N = n;
p = p;
B = rand(M,N) < p;
B=B*(-2)+ones(M,N);
```

In the following, we consider the outliers of perturbed elliptic random matrices.

For any matrix $\mathbf{M}$, we denote the Frobenius norm (or Hilbert–Schmidt norm) $\|\mathbf{M}\|_F$ by the formula

$$\|\mathbf{M}\|_F = \sqrt{\mathrm{Tr}\,(\mathbf{M}\mathbf{M}^H)} = \sqrt{\mathrm{Tr}\,(\mathbf{M}^H\mathbf{M})}$$

We denote the spectral norm by $\|\mathbf{M}\|$.

Definition 6.18.4 (condition C1) Let (ξ_1, ξ_2) be a random vector in $\mathbb{R}^2$, where both ξ_1, ξ_2 have mean zero and unit variance. We set $\rho := \mathbb{E}[\xi_1\xi_2]$. Let $\{X_{ij}\}_{i,j\geqslant 1}$ be an infinite double array of real random variables. For each $n \geq 1$, we define the $n \times n$ random matrix $\mathbf{X}_n = (X_{ij})_{i,j=1}^n$. We say the sequence of random matrices $\{\mathbf{X}_n\}_{n\geqslant 1}$ satisfies condition **C**1 with atom variables (ξ_1, ξ_2) if the following hold:

- $\{Y_{ii} : 1 \leqslant i\} \cup \{(Y_{ij}, Y_{ji}) : 1 \leqslant i \leqslant j\}$ is a collection of independent random elements;
- $\{(Y_{ij}, Y_{ji}) : 1 \leqslant i \leqslant j\}$ is a collection of i.i.d. copies of (ξ_1, ξ_2);
- $\{Y_{ii} : 1 \leqslant i\}$ is a collection of i.i.d. random variables with mean zero and finite variance.

Let $\{\mathbf{X}_n\}_{n\geqslant 1}$ be a sequence of random matrices that satisfy condition **C**1 with atom variables (ξ_1, ξ_2). If $\rho := \mathbb{E}[\xi_1\xi_2] = 1$, then $\{\mathbf{X}_n\}_{n\geqslant 1}$ is a sequence of Wigner real symmetric matrices.

Let ξ be a real random variable with mean zero and unit variance. For each $n \geq 1$, let $\mathbf{X}_n$ be an $n \times n$ matrix whose entries are i.i.d. copies of ξ. Then $\mathbf{X}_n$ is a sequence of random matrices that satisfy condition **C**1. If $\mathbf{X}_n$ is a sequence of random matrices that satisfy condition **C**1, then it was shown in [341] that the limiting empirical spectral density of

$\frac{1}{\sqrt{n}}\mathbf{X}_n$ is given by the uniform distribution on the interior of an ellipse. For $-1 < \rho < 1$, define the ellipsoid

$$\mathcal{E}_\rho := \left\{ z = x + jy \in \mathbb{C} : \frac{x^2}{(1+\rho)^2} + \frac{y^2}{(1-\rho)^2} \leqslant 1 \right\} \tag{6.223}$$

Let

$$F_\rho\left(x, y\right) := \mu_\rho\left(z \in \mathbb{C} : \operatorname{Re}(z) \leqslant x, \operatorname{Im}(z) \leqslant y\right)$$

where μ_ρ is the uniformly probability measure on $\mathcal{E}_\rho$. It will also be convenient to define $\mathcal{E}_\rho$ when $\rho = \pm 1$. For $\rho = 1$, let $\mathcal{E}_1$ be the line segment $[-2, 2]$, and for $\rho = -1$, let $\mathcal{E}_{-1}$ be the line segment $[-2, 2]\sqrt{-1}$ on the imaginary axis[3].

Theorem 6.18.5 Let $\left\{\mathbf{X}_n\right\}_{n\geqslant 1}$ be a sequence of random matrices that satisfies condition **C**1 with atom variables $\left(\xi_1, \xi_2\right)$, where $\rho = \mathbb{E}\left[\xi_1\xi_2\right]$, and assume $-1 < \rho < 1$. For $n \geq 1$, let $\mathbf{A}_n$ be an $n \times n$ matrix, and assume the sequence $\left\{\mathbf{A}_n\right\}_{n\geqslant 1}$ satisfies $\operatorname{rank}\left(\mathbf{A}_n\right) = o(n)$ and

$$\sup_{n\geqslant 1} \frac{1}{n^2} \left\|\mathbf{A}_n\right\|_F^2 < \infty$$

Then the empirical spectral distribution (ESD) of $\frac{1}{\sqrt{n}}\left(\mathbf{X}_n + \mathbf{A}_n\right)$ converges almost surely to $F_\rho\left(x, y\right)$ as $n \to \infty$.

A version of Theorem 6.18.5 holds when ξ_1, ξ_2 are complex random variables [341].

Definition 6.18.6 (Condition C0) Let $\left(\xi_1, \xi_2\right)$ be a random vector in $\mathbb{R}^2$, where both ξ_1, ξ_2 have mean zero and unit variance. We set $\rho := \mathbb{E}\left[\xi_1\xi_2\right]$. For each $n \geq 1$, let $\mathbf{X}_n$ be an $n \times n$ matrix. We say the sequence of random matrices $\left\{\mathbf{X}_n\right\}_{n\geqslant 1}$ satisfies condition **C**0 with atom variables $\left(\xi_1, \xi_2\right)$ if the following conditions hold:

- the sequence $\left\{\mathbf{X}_n\right\}_{n\geqslant 1}$ satisfies condition **C**1 with atom variables $\left(\xi_1, \xi_2\right)$;
- we have

$$M_4 := \max\left\{\mathbb{E}\left|\xi_1\right|^4, \mathbb{E}\left|\xi_2\right|^4\right\} < \infty$$

We define the distance between points $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$ as

$$d\left(\mathbf{x}, \mathbf{y}\right) := \left(\sum_{i=1}^{n}\left(x_i - y_i\right)^2\right)^{1/2} = \left\|\mathbf{x} - \mathbf{y}\right\|$$

which is the Euclidean norm. Let K be a closed and bounded convex set in $\mathbb{R}^n$. For each $\mathbf{x}_0 \in \mathbb{R}^n$ we define the "distance from a point $\mathbf{x}_0$ to a convex set K" by

$$\operatorname{dist}\left(\mathbf{x_0}, K\right) := \min_{\mathbf{x}\in K} d\left(\mathbf{x}_0, \mathbf{x}\right)$$

3 We use $\sqrt{-1}$ to denote the imaginary unit.

Since $\mathcal{E}_\rho$ is a convex set and z is a point $z \in \mathbb{R}^2$, we will define the neighborhoods

$$\mathcal{E}_{\rho,\delta} := \left\{ z \in \mathbb{C} : \operatorname{dist}\left(z, \mathcal{E}_\rho\right) \leqslant \delta \right\}$$

for any $\delta > 0$.

Theorem 6.18.7 Let $\left\{\mathbf{X}_n\right\}_{n\geqslant 1}$ be a sequence of random matrices that satisfies condition $\mathbf{C}0$ with atom variables (ξ_1, ξ_2), where $\rho = \mathbb{E}\left[\xi_1\xi_2\right]$. Let $\delta > 0$. Then, almost surely, for n sufficiently large, all the eigenvalues of $\frac{1}{\sqrt{n}}\mathbf{X}_n$ are contained in $\mathcal{E}_{\rho,\delta}$.

Corollary 6.18.8 (Spectral radius of elliptic random matrices) Let $\left\{\mathbf{X}_n\right\}_{n\geqslant 1}$ be a sequence of random matrices that satisfies condition $\mathbf{C}0$ with atom variables (ξ_1, ξ_2), where $\rho = \mathbb{E}\left[\xi_1\xi_2\right]$. Let $\delta > 0$. Then the spectral radius of $\frac{1}{\sqrt{n}}\mathbf{X}_n$ converges almost surely to $1 + |\rho|$ as $n \to \infty$.

See figures above for illustrations.
Now we are in a position to state the main theorem in this section.

Theorem 6.18.9 (outliers for low rank perturbations of elliptic random matrices) Let $k \geq 1$ and $\delta > 0$. Let $\left\{\mathbf{X}_n\right\}_{n\geqslant 1}$ be a sequence of random matrices that satisfies condition $\mathbf{C}0$ with atom variables (ξ_1, ξ_2), where $\rho = \mathbb{E}\left[\xi_1\xi_2\right]$. For each $n \geq 1$, let $\mathbf{C}_n$ be a deterministic $n \times n$ matrix, where

$$\sup_{n\geqslant 1} \operatorname{rank}\left(\mathbf{C}_n\right) \leqslant k, \text{ and } \sup_{n\geqslant 1} \left\|\mathbf{C}_n\right\| = O(1)$$

Suppose for n sufficiently large, there are no nonzero eigenvalues of $\mathbf{C}_n$ that satisfy

$$\lambda_i\left(\mathbf{C}_n\right) + \frac{\rho}{\lambda_i\left(\mathbf{C}_n\right)} \in \mathcal{E}_{\rho,3\delta} \backslash \mathcal{E}_{\rho,\delta} \text{ with } \left|\lambda_i\left(\mathbf{C}_n\right)\right| > 1$$

and there are j eigenvalues $\lambda_1\left(\mathbf{C}_n\right), \ldots, \lambda_j\left(\mathbf{C}_n\right)$ for some $j \leq k$ that satisfy

$$\lambda_i\left(\mathbf{C}_n\right) + \frac{\rho}{\lambda_i\left(\mathbf{C}_n\right)} \in \mathbb{C} \backslash \mathcal{E}_{\rho,3\delta} \text{ with } \left|\lambda_i\left(\mathbf{C}_n\right)\right| > 1$$

Then, almost surely, for n sufficiently large, there are exactly j eigenvalues of $\frac{1}{\sqrt{n}}\mathbf{X}_n + \mathbf{C}_n$ in the region $\mathbb{C}\backslash\mathcal{E}_{\rho,2\delta}$, and after labeling the eigenvalues properly

$$\lambda_i\left(\frac{1}{\sqrt{n}}\mathbf{X}_n + \mathbf{C}_n\right) = \lambda_i\left(\mathbf{C}_n\right) + \frac{\rho}{\lambda_i\left(\mathbf{C}_n\right)} + o(1)$$

for each $1 \leq i \leq j$.

In [342, 343], spiked deformations of Wigner random matrices plus deterministic matrices are considered. Theorem 6.18.9 can be viewed as a non-Hermitian extension of the results in [342, 343].

We now consider the case of elliptic random matrices with nonzero mean:

$$\frac{1}{\sqrt{n}}\mathbf{X}_n + \mu\sqrt{n}\varphi_n\varphi_n^T$$

where $\{\mathbf{X}_n\}_{n\geqslant 1}$ that satisfies condition **C**0 with atom variables ξ_1, ξ_2, μ is a fixed nonzero complex number (independent of n), and $\varphi_n = \frac{1}{\sqrt{n}}(1, \ldots, 1)^T$. The outliers for elliptic random matrices with nonzero means can be handled. The nonzero mean is a rank one perturbation of $\frac{1}{\sqrt{n}}\mathbf{X}_n$. This corresponds to shifting the entries of $\mathbf{X}_n$ by μ (so they have mean μ instead of mean zero). The elliptic law still holds for this rank one perturbation of $\frac{1}{\sqrt{n}}\mathbf{X}_n$, thanks to Theorem 6.18.5. In view of Theorem 6.18.9, we show there is a single outlier for this ensemble near $\mu\sqrt{n}$.

Theorem 6.18.10 (outlier for elliptic random matrices with nonzero mean) Let $\delta > 0$. Let $\{\mathbf{X}_n\}_{n\geqslant 1}$ be a sequence of random matrices that satisfies condition **C**0 with atom variables (ξ_1, ξ_2), where $\rho = \mathbb{E}[\xi_1\xi_2]$, and let μ be a nonzero complex number independent of n. Then almost surely, for a sufficiently large n, all the eigenvalues of $\frac{1}{\sqrt{n}}\mathbf{X}_n + \mu\sqrt{n}\varphi_n\varphi_n^T$ lie in $\mathcal{E}_{\rho,\delta}$, with a single exception taking the value $\mu\sqrt{n} + o(1)$.

A version of Theorem 2.7 was proven by Furedi and Komlos in [344] for a class of real symmetric Wigner matrices. Moreover, Furedi and Komlos study the fuctuations of the outlier eigenvalue. Tao (2013) [345] verified Theorem 6.18.10 when $\mathbf{X}_n$ is a random matrix with i.i.d. entries.

Bibliographical Remarks

Throughout this chapter we have taken the liberty of borrowing materials from [139]. During his academic visit to NTNU in the summer of 2012, the first author had the privilege of meeting with his thesis adviser Prof. Ralf Müller. The discussions with Prof. Ralf Müller were very useful. Relevant work includes [346–348].

The results in Section 6.8 can be found in [139] and [349, 350].

In Section 6.8 we draw on material from [274] and [293].

We followed [311, 312] in Section 6.2.4.

Section 6.8, we followed [309] for part of our exposition. In Section 6.9, we [73, 351] for part of our exposition.

Section 6.10 is taken from [75].

In Section 6.11, we followed [311, 312].

Section 6.12 primarily followed [352, 353]. In [354] Akemann, Ipsen and Kieburg discuss the product of L *rectangular* random matrices with independent Gaussian entries. In [287], Ipsen and Kieburg study the joint probability density of the eigenvalues of a product of rectangular real, complex or quaternion random matrices in a unified way. The random matrices are distributed according to arbitrary probability densities, whose only restriction is the invariance under left and right multiplication by orthogonal, unitary or unitary symplectic matrices, respectively. They show that a product of rectangular matrices is statistically equivalent to a product of square matrices. In this way they prove a weak commutation relation of the random matrices at finite matrix sizes, which have previously been discussed for infinite matrix size. In [355], Forrester studied the probability that all eigenvalues are real for the matrix product $\mathbf{P}_L = \mathbf{X}_L\mathbf{X}_{L-1}\cdots\mathbf{X}_1$, where $\mathbf{X}_i, i = 1, \ldots, L$ independent $N \times N$ standard Gaussian random matrices.

In Section 6.13, we follow [315, 318, 319, 356, 357] for our exposition.

In Section 6.14, our interest in Euclidean random matrices is inspired by their applications in physics [320, 321, 358–361]. The relevant mathematical literature includes [362–366]. In high dimension space, concentration of measure phenomenon [40] naturally occurs. Both Hermitian and Non-Hermitian Euclidean random matrices can be studied. The theory is applied to the random Green's matrix relevant to wave propagation in an ensemble of point-like scattering centers [321, 358]. Wave propagation in random media is directly relevant to massive MIMO, a disruptive technology in 5G wireless technology. We closely follow [321], [320] and [358, 359] for the exposition.

In Section 6.14, despite the significance of the random Green's matrix (6.157), little is known about statistical properties of its eigenvalues complex and their probability distribution is difficult to access. The principal difficulties that one encounters when trying to develop a theory of non-Hermitian Euclidean random matrices stem from the nontrivial statistics of their elements and the correlations between them. Both are not known analytically and are often difficult to calculate. The first paper for an analytical theory is [321]. Numerical simulations are typically used. Some analytic results are available in the limit of high density of points $\mathbf{r}_i$ inside a sphere: $\rho = N/V \to \infty$, when the summation in the eigenvalue equation $\sum_j G_{ij}\psi_i = \lambda_i\psi_i$ can be replaced by integration. The work of Skipetrov and Goetschy (2011) [320], partially fills this gap by considering eigenvalue distributions of the three matrices $\mathbf{G}, \mathbf{S} = \mathrm{Im}\,\mathbf{G}, \mathbf{C} = \mathrm{Re}\,\mathbf{G}$ at *finite* densities ρ, with the distances between neighboring points $\mathbf{r}_i$ that are larger than, comparable to, or smaller than the wavelength $\lambda_0 = 2\pi/k_0$. This situation is of particular importance in the context of wave propagation in random media because in order to observe phenomena due to scattering of waves on the heterogeneities of the medium, the density of scattering centers (or scatterers) should be neither too low (in this case the scattering is negligible), nor too high (in this case the medium responds as an effective homogeneous medium).

This line of research introduced in Section 6.14 deserves further investigation in the context of massive MIMO, with applications in both communications and sensing/radar. Waveforms for modulation or sensing may be designed by using the metrics in terms of eigenvalue distributions $p(\lambda)$. This has been made possible by the analytical machinery introduced in Section 6.14.

In Section 8.8, we will study how large random matrices are perturbed by finite rank perturbations, drawing material from [367].

In Section 6.18, we take material from [340]. The discovery of outliers in Figure 6.39 was made in July 2013 during the writing of Section 6.18: See Figure 6.39 and the MATLAB code "Outliers in the Elliptic Law." The paper by O'Rourke and Renfrew (2013) [368] was posted in September 2013. The outlier part of Section 6.18 is taken from this. In Section 6.15, we drew material from [328, 329], [340] and [368].

In Section 6.16, outliers in the spectrum of i.i.d. matrices with bounded rank perturbations were considered, taking material from [345]. Another good reference is [328].

In Section 6.17, we took the liberty of freely drawing material from [138].

Some parts of our exposition in Section 6.6 followed [302].

7

The Mathematical Foundations of Data Collection

The previous chapters serve as the mathematical foundation for big data representation and analytics. We assume that massive datasets are available. This topic is very important. We have postponed this topic until now because the necessary mathematical foundation was not complete. The covariance matrix for big data is of central interest. We will review methods for covariance matrix estimation. A sample covariance matrix, a Hermitian positive non-negative random matrix, is traditionally used. This subject needs to be revisited when the dimensions of data matrices are large. This subject is naturally connected with compressive sensing.

For big data, we must rethink traditional data collection. We must focus on the global picture, considering the collection, storage, cleaning and processing of data as a whole. Among other things, a central task is to offer a *principled* and *automated* way of selecting regularization parameters in a variety of problems. The most important legacy of compressive sensing may be that it has forced us to think about information, complexity, hardware, and algorithms in a truly integrated manner. Compressive sensing can be viewed as a good example of applications for large random matrices [40].

Data storage is central to big data. Real-time processing is demanding. For many applications, we often cannot afford the luxury of saving all the raw data generated by the system (or network) for future processing. One fundamental challenge is to choose what types of information are stored. As we are dealing with streaming data, real-time processing is required.

Large random matrices are the unifying theme of this chapter. This starts with data storage using large random matrices. High throughput real-time processing is required.

As pointed out in Section 1.1.5, it may be necessary to rethink data collection and storage to facilitate big-data processing and inference tasks. How do we trade-off complexity for accuracy in massive decentralized signal and data-analysis tasks? What are the basic principles and useful methodologies to scale inference and learning algorithms and trade off the computational resources (e.g., time, space and energy) according to the needs of engineering practice (e.g. robustness versus real-time efficiency)?

7.1 Architectures and Applications for Big Data

Significant topics include:

- Scalable, distributed computing, for example MapReduce, Hadoop.
- Streaming for real-time analytics and graph processing, for example Pregel, Giraph.

Smart Grid using Big Data Analytics: A Random Matrix Theory Approach, First Edition.
Robert C. Qiu and Paul Antonik.

- Smart power-grid analytics.
- Multimodal sensing.
- Preference measurement; recommender systems; targeted advertising.
- Data collection, storage and transmission.
- Sampling.

In this chapter, we use the large random matrix as a single tool to explore these topics. Streaming for real-time analytics is the aim of this chapter. Preference measurement can be formulated in terms of covariance matrix estimation. We emphasize abstract statistical models rather than specific applications so that the principles may be applied more widely. In the infancy of big data, this philosophy is justified.

7.2 Covariance Matrix Estimation

High-dimensional covariance estimation is known to be a difficult problem, in the"large p small n" setting. Our goal is to collect data to obtain the estimator $\hat{\mathbf{\Sigma}}$ for the true covariance matrix $\mathbf{\Sigma}$. During real-time data collection of streaming data, we sometimes encounter the "large p small n," to minimize the delay in data collection. In recent years, the availability of high-throughput data from various applications has pushed this problem to an extreme where, in many situations, the number of samples (n) is often much smaller than the number of parameters (p). When $n < p$, the sample covariance matrix $\mathbf{S}$ is singular and not positive definite and hence it cannot be inverted to compute the precision matrix (the inverse of the covariance matrix). However, even when $n > p$, the eigenstructure tends to be systematically distorted unless p/n is extremely small, resulting in ill-conditioned estimators for $\mathbf{\Sigma}$; see [369] and [370].

Since the seminal work in [369] and [370] the problem of estimating $\mathbf{\Sigma}$ has been recognized as highly challenging. Formally, given n independent sample vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ from a zero-mean p-dimensional Gaussian distribution with an unknown covariance matrix $\mathbf{\Sigma}$, the log-likelihood function of the covariance matrix has the form

$$\begin{aligned} L(\mathbf{\Sigma}) &= \log \prod_{i=1}^{n} \frac{1}{(2\pi)^p |\mathbf{\Sigma}|} \exp\left(-\frac{1}{2}\mathbf{x}_i^T \mathbf{\Sigma}^{-1} \mathbf{x}_i\right) \\ &= -(np/2)\log(2\pi) - (n/2)\left(\mathrm{Tr}\left[\mathbf{\Sigma}^{-1}\mathbf{S}\right]\right) - \log\det\mathbf{\Sigma}^{-1} \end{aligned}$$

where both $|\mathbf{\Sigma}|$ and $\det \mathbf{\Sigma}$ denote the determinant of $\mathbf{\Sigma}$, $\mathrm{Tr}(\mathbf{A})$ denotes the trace of $\mathbf{A}$, and $\mathbf{S}$ is the sample covariance matrix, i.e.,

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$$

The log-likelihood function is maximized by the sample covariance, i.e., the maximum likelihood estimate (MLE) of the covariance is $\mathbf{S}$ [371].

The negative log-likelihood function of $\mathbf{\Sigma}$ given the sample, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, is proportional to

$$L_n(\mathbf{\Sigma}) = -\log\det\mathbf{\Sigma}^{-1} + \mathrm{Tr}\left[\mathbf{\Sigma}^{-1}\mathbf{S}\right] \tag{7.1}$$

up to some constant. When $p < n$, $\mathbf{S}$ is the maximum likelihood estimate of $\mathbf{\Sigma}$. It is well known that $\mathbf{S}$ is not a stable estimate of $\mathbf{\Sigma}$ when p is large or p is close to the sample size n. As the dimension p increases, the largest eigenvalues of $\mathbf{S}$ tend to be systematically distorted, which can give an ill-conditioned estimate of $\mathbf{\Sigma}$ [177]. When $p > n$, $\mathbf{S}$ is singular and the smallest eigenvalue is zero. It is not appropriate to use $\mathbf{S}$ to obtain the estimate of $\mathbf{\Sigma}^{-1}$.

Research is done to explore better alternative estimators for $\mathbf{\Sigma}$ (or $\mathbf{\Sigma}^{-1}$) in both the frequentist and Bayesian frameworks. Many of these estimators give substantial risk reductions compared to the sample covariance estimator $\mathbf{S}$ in small sample sizes. A common underlying property of many of these estimators is that they are shrinkage estimators in the James–Stein sense [372] and [373]). Warton [374] minimizes the predictive risk, which is estimated using a crossvalidation method. Many other James–Stein type shrinkage estimators have been studied from a decision-theoretic point of view.

A simple example is a family of linear shrinkage estimators, which take a convex combination of the sample covariance and a suitably chosen target or regularization matrix. Ledoit and Wolf [375] studied a linear shrinkage estimator towards a specified target covariance matrix, and chose the optimal shrinkage to minimize the Frobenius risk.

Regularized likelihood methods for the multivariate Gaussian model provide estimators with different types of shrinkage. Sheenaand Gupta [376] propose a constrained maximum-likelihood estimator with constraints on the smallest or the largest eigenvalues. Generally speaking, the feasibility of using convex optimization for some real-time computation opens the door for many applications, including data collection.

To take advantage of some prior information about $\mathbf{\Sigma}$, we can use the loss functions to estimate the true covariance matrix. We examine some properties of these loss functions, recalled from Section 8.9.2 for convenience.

$$L\left(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}\right) = \operatorname{Tr} \mathbf{\Sigma}^{-1}\hat{\mathbf{\Sigma}} - \log\det \mathbf{\Sigma}^{-1}\hat{\mathbf{\Sigma}} - p \tag{7.2}$$

We will use the loss function defined by (8.65) largely because it is comparatively easy to work with this loss function. However, it also has all the appealing properties of loss functions:

1) $L\left(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}\right) \geqslant 0$, with equality if and only if $\mathbf{\Sigma} = \hat{\mathbf{\Sigma}}$.
2) $L\left(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}\right)$ a convex function of its second argument;
3) $L\left(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}\right)$ invariant under linear transformations of $\mathbb{R}^n$, i.e., for any nonsingular $p \times p$ matrix $\mathbf{A}$,

$$L\left(\mathbf{A}\mathbf{\Sigma}\mathbf{A}^H, \mathbf{A}\hat{\mathbf{\Sigma}}\mathbf{A}^H\right) = L\left(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}\right) \tag{7.3}$$

We can formulate the data collection problem in terms of convex optimization:

$$\begin{aligned} &\text{minimize} \quad L\left(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}\right) \\ &\text{subject to} \quad \mathrm{F}_i\left(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}\right) = 1, \quad i = 1, \ldots, N \end{aligned} \tag{7.4}$$

where $\mathrm{F}_i\left(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}\right), i = 1, \ldots, N$ are convex functions that give constraints for data collection. Note that $L\left(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}\right)$ is a convex function of the second argument $\hat{\mathbf{\Sigma}}$.

Once a problem is formulated in terms of convex optimization, solving for the problem can call on a standard software solver such as CVX. Thus, (7.4) may be efficiently solved using standard methods such as interior-point methods [377] when the number of variables (i.e., entries in the matrix) is modest, say, under 1000. As the number of variables is about $p(p+1)/2$, the limit is around $p = 45$. In particular, we may use convex optimizations toolboxes such as the CVX in MATLAB [378] and the CVXOPT in Python programming language [379]. The challenge arises from the fact that the convex optimization must be solved in real time manner for streaming data.

Consider an estimation with a condition number constraint. The condition number of a positive definite matrix $\mathbf{A} \geq 0$ is defined as

$$\text{cond}\,(\mathbf{A}) = \lambda_{\max}(\mathbf{\Sigma}) / \lambda_{\min}(\mathbf{A})$$

where $\lambda_{\max}(\mathbf{\Sigma})$ and $\lambda_{\min}(\mathbf{\Sigma})$ are the maximum and the minimum eigenvalues of $\mathbf{A}$, respectively. In several applications a stable well-conditioned estimate of the covariance matrix is required. In other words, we require

$$\text{cond}\,(\mathbf{\Sigma}) \leqslant \kappa_{\max}$$

for a given threshold $\kappa_{\max}$.

The maximum likelihood estimation problem with the condition number constraint can be formulated as

$$\begin{aligned} &\text{maximize } L(\mathbf{\Sigma}) \\ &\text{subject to } \lambda_{\max}(\mathbf{\Sigma}) / \lambda_{\min}(\mathbf{\Sigma}) \leqslant \kappa_{\max} \end{aligned} \tag{7.5}$$

An implicit condition is that $\mathbf{\Sigma}$ is symmetric and positive definite. This problem is a generalization of the problem considered in [376], where only either the lower bound or the upper bound is considered. The covariance estimation problem (7.5) can be reformulated as a convex optimization problem.

Let us use another concrete example of matrix log-transformation to illustrate how data collection is formulated in terms of convex optimization. Consider the spectral decomposition of the covariance matrix $\mathbf{\Sigma} = \mathbf{U}\mathbf{D}\mathbf{U}^T$, where $\mathbf{D} = \text{diag}\left(d_1, \ldots, d_p\right)$ is a diagonal matrix of the eigenvalues of $\mathbf{\Sigma}$, and $\mathbf{U}$ is an orthonormal matrix consisting of eigenvectors of $\mathbf{\Sigma}$. Assume that $d_1 \geqslant d_2 \geqslant \cdots \geqslant d_p \geqslant 0$. Let

$$\mathbf{A} = \left(a_{ij}\right)_{p \times p} = \log(\mathbf{\Sigma})$$

be the matrix logarithm of $\mathbf{\Sigma}$. That is

$$\mathbf{\Sigma} = \sum_{k=0}^{\infty} \frac{1}{k!} \mathbf{A}^k \equiv \exp(\mathbf{A})$$

where $\exp(\mathbf{A})$ is called the matrix exponential of $\mathbf{A}$. Then

$$\mathbf{A} = \mathbf{U}\,\text{diag}\left(\log\left(d_1\right), \ldots, \log\left(d_p\right)\right)\mathbf{U}^T \equiv \mathbf{U}\mathbf{M}\mathbf{U}^T \tag{7.6}$$

where $\mathbf{M}$ is a diagonal matrix. In terms of $\mathbf{A}$, the negative log-likelihood function in (1.1) becomes

$$L_n(\mathbf{A}) = \text{Tr}(\mathbf{A}) + \text{Tr}\left[\exp(-\mathbf{A})\,\mathbf{S}\right] \tag{7.7}$$

A major advantage of using the matrix logarithm transformation is that it converts the problem of estimating a *positive definite* matrix $\mathbf{\Sigma}$ into a problem of estimating a *real symmetric* matrix A. Using the Volterra integral equation [380], we have

$$\exp(\mathbf{A}t) = \exp\left(\mathbf{A}_0 t\right) + \int_0^t \exp\left(\mathbf{A}_0 (t-s)\right)\left(\mathbf{A} - \mathbf{A}_0\right)\exp(\mathbf{A}s)\, ds, \quad 0 < t < \infty \tag{7.8}$$

Let $\mathbf{\Sigma}_0$ be an initial estimate of $\mathbf{\Sigma}$ and $\mathbf{A}_0 = \log\left(\mathbf{\Sigma}_0\right)$. Using (7.8), and after some manipulation, we approximate $L_n(\mathbf{A})$

$$\begin{aligned}\ell_n(\mathbf{A}) = \operatorname{Tr}\left[\mathbf{\Sigma}_0^{-1}\mathbf{S}\right] - \left[\int_0^1 \operatorname{Tr}\left[\left(\mathbf{A} - \mathbf{A}_0\right)\mathbf{\Sigma}_0^{-s}\mathbf{S}\mathbf{\Sigma}_0^{s-1}\right] ds - \operatorname{Tr}(\mathbf{A})\right] \\ + \int_0^1 \int_0^s \operatorname{Tr}\left[\left(\mathbf{A} - \mathbf{A}_0\right)\mathbf{\Sigma}_0^{u-s}\left(\mathbf{A} - \mathbf{A}_0\right)\mathbf{\Sigma}_0^{-u}\mathbf{S}\mathbf{\Sigma}_0^{s-1}\right] du ds\end{aligned} \tag{7.9}$$

The integrations in (7.9) can be analytically solved through the spectral decomposition $\mathbf{\Sigma}_0 = \mathbf{U}_0\mathbf{D}_0\mathbf{U}_0^T$. Define

$$\begin{aligned}&\mathbf{B} = \mathbf{U}_0^T\left(\mathbf{A} - \mathbf{A}_0\right)\mathbf{U}_0 = \left(b_{ij}\right)_{p\times p}, \\ &\tilde{\mathbf{S}} = \mathbf{U}_0^T\mathbf{S}\mathbf{U}_0 = \left(\tilde{s}_{ij}\right)_{p\times p}, \mathbf{D}_0 = \operatorname{diag}\left(d_1^{(0)}, \ldots, d_p^{(0)}\right)\end{aligned}$$

We obtain

$$\begin{aligned}\ell_n(\mathbf{A}) = \sum_{i=1}^p \frac{1}{2}\xi_{ii}b_{ii}^2 + \sum_{i<j}\xi_{ij}b_{ij}^2 + 2\sum_{i=1}^p\sum_{j\neq i}\tau_{ij}b_{ii}b_{ij} + \sum_{i=1}^p \sum_{i<j, i\neq k, j\neq k}\eta_{kij}b_{ik}b_{kj} \\ - \left[\sum_{i=1}^p \beta_{ii}b_{ii} + 2\sum_{i<j}\beta_{ij}b_{ij}\right]\end{aligned} \tag{7.10}$$

up to some constant. We see that $\ell_n(\mathbf{A})$ in (7.10) is a quadratic function of b_{ij}. As the matrix $\mathbf{B}$ is a linear transformation of $\mathbf{A}$, $\ell_n(\mathbf{A})$ is also a quadratic function of $\mathbf{A}$. The coefficients in (7.10) are functions of $\left(\tilde{s}_{ij}\right)_{p\times p}$ and $d_1^{(0)}, \ldots, d_p^{(0)}$.

We apply a regularized approach to estimating $\mathbf{\Sigma}$ by using the approximate log-likelihood function $\ell_n(\mathbf{A})$ in (7.10). Consider the penalty function $\|\mathbf{A}\|_F^2$, the Frobenius norm of $\mathbf{A}$, which is equivalent to $\operatorname{Tr}\left(\mathbf{A}^2\right)$. From (7.11)

$$\operatorname{Tr}\left(\mathbf{A}^2\right) = \sum_{i=1}^p \left(\log\left(d_i\right)\right)^2$$

where d_i is the eigenvalue of the covariance matrix $\mathbf{\Sigma}$. If d_i goes to zero or diverges to infinity, the value of $\log(d_i)$ goes to infinity in both cases. Therefore, such a penalty function can simultaneously regularize the largest and smallest eigenvalues of the covariance matrix estimate. We consider to estimate $\mathbf{\Sigma}$, or equivalently $\mathbf{A}$, by minimizing

$$\ell_{n,\lambda}(\mathbf{A}) = \ell_n(\mathbf{A}) + \lambda\operatorname{Tr}\left(\mathbf{A}^2\right) \tag{7.11}$$

where λ is a tuning parameter. Note that $\operatorname{Tr}\left(\mathbf{A}^2\right) = \operatorname{Tr}\left(\mathbf{U}_0\mathbf{B}\mathbf{U}_0^T + \mathbf{A}_0\right)^2$ is equivalent to $\operatorname{Tr}\left(\mathbf{B}^2\right) + 2\operatorname{Tr}(\mathbf{B}\mathbf{\Gamma})$ up to some constant, where $\mathbf{\Gamma} = \left(\gamma_{ij}\right)_{p\times p} = \mathbf{U}_0^T\mathbf{A}_0\mathbf{U}_0$. Then (7.11) becomes

$$\ell_{n,\lambda}(\mathbf{B}) = \sum_{i=1}^{p} \frac{1}{2}\xi_{ii} b_{ii}^2 + \sum_{i<j} \xi_{ij} b_{ij}^2 + 2\sum_{i=1}^{p}\sum_{j\neq i} \tau_{ij} b_{ii} b_{ij} + \sum_{i=1}^{p}\sum_{i<j, i\neq k, j\neq k} \eta_{kij} b_{ik} b_{kj} - \left[\sum_{i=1}^{p} \beta_{ii} b_{ii} + 2\sum_{i<j} \beta_{ij} b_{ij}\right] + \lambda\left[\sum_{i=1}^{p} \frac{1}{2}\xi_{ii} b_{ii}^2 + \sum_{i=1}^{p} b_{ij}^2 + \sum_{i<j} \gamma_{ii} b_{ii} + \sum_{i<j} \gamma_{ij} b_{ij}\right] \tag{7.12}$$

Let $\hat{\mathbf{B}}$ be the minimizer of (7.12). The iterative algorithm is described as follows:

1) Set an initial covariance matrix estimate $\mathbf{\Sigma}_0$, a positive definite matrix.
2) Obtain the spectral decomposition $\mathbf{\Sigma}_0 = \mathbf{U}_0\mathbf{D}_0\mathbf{U}_0^T$, and set $\mathbf{A}_0 = \log\left(\mathbf{\Sigma}_0\right)$.
3) Compute $\hat{\mathbf{B}}$ by minimizing $\ell_{n,\lambda}(\mathbf{B})$ in (7.12). Then obtain $\hat{\mathbf{A}} = \mathbf{U}_0\hat{\mathbf{B}}\mathbf{U}_0^T + \mathbf{A}_0$, update the estimate of $\mathbf{\Sigma}$ by
$$\hat{\mathbf{\Sigma}} = \exp\left(\hat{\mathbf{A}}\right) = \exp\left(\mathbf{U}_0\hat{\mathbf{B}}\mathbf{U}_0^T + \mathbf{A}_0\right)$$
4) Check if $\left\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_F^2$ is less than a prespecified positive tolerance value. Otherwise, set $\mathbf{\Sigma}_0 = \hat{\mathbf{\Sigma}}$ and go back to Step 2.

The performance of this iterative algorithm is the best among the state-of-the-art algorithms with which it has been compared, including the sample covariance matrix, which is the maximum likelihood estimator. It is also better than the maximum likelihood covariance estimator with a condition number constraint. See [381] for details.

7.3 Spectral Estimators for Large Random Matrices

Only the spectrum properties (eigenvalues) of large random matrices can be stored for future data processing. For an $n \times n$ matrix $\mathbf{X}$, instead of storing the n^2 entries of $\mathbf{X}$, we store the n eigenvalues of $\mathbf{X}$. We reduce the dimensionality by n times. When n is large, for instance $n = 10^3$, the saving of the required storage is significant (1000 times). Real-time processing includes estimating covariance matrix and calculating eigenvalues. Principal component analysis (PCA) is a well established dimensionality reduction method commonly used to denoise and visualize data.

Our problem is to recover an approximately low-rank data matrix from noisy observations. We introduce an unbiased risk estimate—holding in a Gaussian model—for any spectral estimator obeying some mild regularity assumptions. In particular, we give an unbiased risk estimate formula for singular-value thresholding (SVT), a popular estimation strategy that applies a soft-thresholding rule to the singular values of the noisy observations. Among other things, our formulas offer a principled and automated way of selecting regularization parameters in a variety of problems.

Suppose we have noisy observation matrix $\mathbf{Y}$ about an $m \times n$ data matrix $\mathbf{X}_0$ of interest:

$$\mathbf{Y} = \mathbf{X}_0 + \mathbf{Z}, \qquad \text{where } Z_{ij} \sim \text{i.i.d. } \mathcal{N}(0, 1) \tag{7.13}$$

We wish to estimate $\mathbf{X}_0$ as accurately as possible. For our problem at hand, the estimation process—part of data collection—must be done in real time. Our motivation is to reduce dimensionality. The estimand has some structure, namely $\mathbf{X}_0$ has low rank or is well approximated by a low-rank matrix. This assumption is often met in practice as the columns of $\mathbf{X}_0$ can be quite correlated.

7.3.1 Singular Value Thresholding

Whenever the object of interest has (approximately) low rank, it is possible to improve upon the naive estimate $\hat{\mathbf{X}}_0 = \mathbf{Y}$ by regularizing the maximum likelihood. A natural approach consists in truncating the singular value decomposition of the observed matrix $\mathbf{Y}$, and solve

$$\mathrm{SVHT}_\lambda(\mathbf{Y}) = \underset{\mathbf{X}\in\mathbb{R}^{m\times n}}{\arg\min}\ \frac{1}{2}\|\mathbf{Y}-\mathbf{X}\|_F^2 + \lambda\,\mathrm{rank}\,(\mathbf{X}) \tag{7.14}$$

where λ a positive scalar. If

$$\mathbf{Y} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^H = \sum_{i=1}^{\min(m,n)} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \tag{7.15}$$

is a singular value decomposition for $\mathbf{Y}$, the solution is given by retaining only the part of the expansion with singular values exceeding the threshold λ:

$$\mathrm{SVHT}_\lambda(\mathbf{Y}) = \sum_{i=1}^{\min(m,n)} \mathbb{I}\left(\sigma_i > \lambda\right) \mathbf{u}_i \mathbf{v}_i^T$$

where $\mathbb{I}$ is the indicator function of the set. In other words, one applies a hard-thresholding rule to the singular values of the observed matrix $\mathbf{Y}$. Such an estimator is discontinuous in $\mathbf{Y}$ and a popular alternative approach applies, instead, a soft-thresholding rule to the singular values:

$$\mathrm{SVST}_\lambda(\mathbf{Y}) = \sum_{i=1}^{\min(m,n)} \left(\sigma_i - \lambda\right)_+ \mathbf{u}_i \mathbf{v}_i^T \tag{7.16}$$

In other words, we shrink the singular values towards zero by a constant amount λ. The estimate $\mathrm{SVST}_\lambda(\mathbf{Y})$ is a *Lipschitz-continuous* function. This follows from the fact that the singular value thresholding operation (7.16) is the prox of the nuclear norm $||\cdot||_*$ (the nuclear norm of a matrix is sum of its singular values); it is the unique solution to

$$\text{minimize} \quad \frac{1}{2}\|\mathbf{Y}-\mathbf{X}\|_F^2 + \lambda\|\mathbf{X}\|_* \tag{7.17}$$

Let $g : \mathbb{R}^M \to \mathbb{R}^N$: we regard the $n \times n$ (respectively, symmetric) matrix space as a special case of $\mathbb{R}^N$ with $M = n^2$ (respectively, $M = n(n+1)/2$). Hence the discussions here apply to matrix variable and/or matrix valued functions as well. Let $||\cdot||$ denote the ℓ_2 norm in finite dimensional Euclidean spaces. Recall that $g(\mathbf{x})$ is said to be locally Lipschitz continuous around $\mathbf{x} \in \mathbb{R}^M$ if there exist a constant κ and an open neighborhood $\mathcal{N}$ of $\mathbf{x}$ such that

$$\|g(\mathbf{y}) - g(\mathbf{z})\| \leqslant \kappa\,\|\mathbf{y}-\mathbf{z}\| \quad \forall \mathbf{y}, \mathbf{z} \in \mathcal{N}$$

We call g a locally Lipschitz function if it is locally Lipschitz continuous around every point of $\mathbb{R}^M$. Moreover, if the above inequality holds for $\mathcal{N} = \mathbb{R}^M$, then g is said to be globally Lipschitz continuous with Lipschitz constant κ.

7.3.2 Stein's Unbiased Risk Estimate (SURE)

A function $g : \mathbb{R}^n \mapsto \mathbb{R}$ is said to be *weakly differentiable* with respect to the variable x_i if there exists $h : \mathbb{R}^n \mapsto \mathbb{R}$ such that for all compactly supported and infinitely differentiable functions ϕ

$$\int \varphi(\mathbf{x}) h(\mathbf{x})\, d\mathbf{x} = -\int \frac{\partial \varphi(\mathbf{x})}{\partial x_i} g(\mathbf{x})\, d\mathbf{x}$$

where $\mathbf{x} = \left(x_1, \ldots, x_p\right)^T$.

Stein [382] gave a formula for an unbiased estimate of the mean-squared error of an estimator obeying a weak differentiability assumption and mild integrability conditions. Roughly speaking, the derivatives can fail to exist over regions of Lebesgue measure zero.

Proposition 7.3.1 (Stein (1981) [382] and Johnstone (2007) [383]) Suppose that $Y_{ij} \sim i.i.d \quad \mathcal{N}\left(X_{ij}, 1\right)$. Consider an estimator $\hat{\mathbf{X}}$ of the form $\mathbf{X} = \mathbf{Y} + g(\mathbf{Y})$, where $g_{ij} : \mathbb{R}^{m \times m} \mapsto \mathbb{R}$ is weakly differentiable with respect to Y_{ij} and

$$\mathbb{E}\left\{ \left| Y_{ij} g_{ij}(\mathbf{Y}) \right| + \left| \frac{\partial}{\partial Y_{ij}} g_{ij}(\mathbf{Y}) \right| \right\} < \infty$$

for $\left(i, j\right) \in \mathcal{I} := \{1, \ldots, m\} \times \{1, \ldots, n\}$. Then

$$\mathbb{E}\left\| \hat{\mathbf{X}} - \mathbf{X} \right\|_F^2 = \mathbb{E}\left\{ mn + 2\,\mathrm{div}\left(g(\mathbf{Y})\right) + \|g(\mathbf{Y})\|_F^2 \right\} \tag{7.18}$$

It was established in [384] that SVST obeys these assumptions, and the deduction of a closed-form expression for its divergence was obtained by the authors.

The classical question is how much shrinkage should be applied. Too much shrinkage gives a large bias whereas too little results in a high variance. To find the correct tradeoff, it would be desirable to have a method that would allow us to compare the quality of estimation for different values of the parameter λ. Ideally, we would like to select λ as to minimize the mean-squared error or risk

$$\mathrm{MSE}(\lambda) = \mathbb{E}\left\| \mathbf{X}_0 - \mathrm{SVST}_\lambda(\mathbf{Y}) \right\|_F^2 \tag{7.19}$$

This cannot be achieved, however, because the expectation in (7.19) depends on the true $\mathbf{X}_0$, and is thus unknown. When the observations follow the model (7.13), it is possible to construct an unbiased estimate of the risk, namely, Stein's unbiased risk estimate (SURE) [382] given by

$$\mathrm{SURE}\left(\mathrm{SVST}_\lambda\right)(\mathbf{Y}) = -mn\tau^2 + \sum_{i=1}^{\min(m,n)} \min\left(\lambda^2, \sigma_i^2\right) + 2\tau^2\,\mathrm{div}\left(\mathrm{SVST}_\lambda(\mathbf{Y})\right) \tag{7.20}$$

where $\left\{\sigma_i\right\}_{i=1}^n$ denotes the singular values of $\mathbf{Y}$. Here, 'div' is the divergence of the nonlinear mapping SVST_λ, which is to be interpreted in a weak sense. Roughly speaking, it can fail to exist on negligible sets.

The main contribution of [384] is to provide a closed-form expression for the divergence of this estimator. They prove that, in the real-valued case

$$\operatorname{div}\left(\mathrm{SVST}_{\lambda}(\mathbf{Y})\right) = \sum_{i=1}^{\min(m,n)} \left[\mathbb{I}(\sigma_i > \lambda) + |m-n|\left(1-\frac{\lambda}{\sigma_i}\right)_+\right] + 2\sum_{i\neq j, i,j=1}^{\min(m,n)} \frac{\sigma_i(\sigma_i-\lambda)_+}{\sigma_i^2-\sigma_j^2} \tag{7.21}$$

when $\mathbf{Y}$ is simple—it has no repeated singular values—and 0 otherwise, say, is a valid expression for the weak divergence. Hence, this formula can be used in (7.20), and gives the determination of a suitable threshold level by minimizing the estimate of the risk, which only depends upon the observed data.

Example 7.3.2 (MATLAB experiments) We work with four matrices $\mathbf{X}_0^{(i)}, i = 1, \ldots, 4$ of size 200×500. Here, $\mathbf{X}_0^{(1)}$ has full rank; $\mathbf{X}_0^{(2)}$ has rank 100; $\mathbf{X}_0^{(3)}$ has rank 10; and $\mathbf{X}_0^{(4)}$ has singular values equal to $\sigma_i = \sqrt{200}/\left(1+e^{(i-100)/20}\right), i = 1, \ldots, 200$. Each matrix is normalized so that $\left\|\mathbf{X}_0^{(i)}\right\|_F = 1, i = 1, \ldots, 4$. Next, two methods are used to estimate the risk (7.20) of SVST_{λ} seen as a function of λ. The first methods uses

$$\hat{R}_i(\lambda) = \frac{1}{N}\sum_{j=1}^{N}\left\|\mathrm{SVST}_{\lambda}\left(\mathbf{Y}_j^{(i)}\right) - \mathbf{X}_0^{(i)}\right\|_F^2 \tag{7.22}$$

where $\left\{\mathbf{Y}_j^{(i)}\right\}_{j=1}^{N}, i = 1, \ldots, 4, N = 50$ are independent samples drawn from model (7.13) with $\mathbf{X}_0 = \mathbf{X}_0^{(i)}, i = 1, \ldots, 4$. The second uses $\mathrm{SURE}\left(\mathrm{SVST}_{\lambda}\right)(\mathbf{Y})$, where $\mathbf{Y}$ is drawn from model (7.13) from $\left\{\mathbf{Y}_j^{(i)}\right\}_{j=1}^{N}, i = 1, \ldots, 4$. Finally, in each case we use values of the signal-to-noise ratio, defined as $\mathrm{SNR} = \left\|\mathbf{X}_0^{(i)}\right\|_F/\sqrt{mn}\tau = 1/\sqrt{mn}\tau$, and set $\mathrm{SNR} = 0.5, 1, 2, 4$. As shown in Figure 7.1 and Figure 7.2, SURE remains very close to the true value of the risk, even though it is calculated from a single observation. Matlab code reproducing the figures is available and computing SURE formulas for various spectral estimators are available in [385]. □

Observations can take on complex values. The model (7.13) has to be modified as

$$\mathbf{Y} = \mathbf{X}_0 + \mathbf{Z}, \qquad \text{where } \mathrm{Re}\left(Z_{ij}\right), \mathrm{Im}\left(Z_{ij}\right) \sim \text{i.i.d. } \mathcal{N}(0,1) \tag{7.23}$$

where the real and imaginary parts are also independent. In this case, SURE becomes

$$\mathrm{SURE}\left(\mathrm{SVST}_{\lambda}\right)(\mathbf{Y}) = -2mn\tau^2 + \sum_{i=1}^{\min(m,n)} \min\left(\lambda^2, \sigma_i^2\right) + 2\tau^2 \operatorname{div}\left(\mathrm{SVST}_{\lambda}(\mathbf{Y})\right) \tag{7.24}$$

We also provide an expression for the weak divergence in this context, namely

$$\begin{aligned}\operatorname{div}\left(\mathrm{SVST}_{\lambda}(\mathbf{Y})\right) = &\sum_{i=1}^{\min(m,n)}\left[\mathbb{I}\left(\sigma_i > \lambda\right) + \left(2|m-n|+1\right)\left(1-\frac{\lambda}{\sigma_i}\right)_+\right] \\ &+4\sum_{i\neq j, i,j=1}^{\min(m,n)} \frac{\sigma_i\left(\sigma_i-\lambda\right)_+}{\sigma_i^2-\sigma_j^2}\end{aligned} \tag{7.25}$$

when $\mathbf{Y}$ is simple and 0 otherwise.

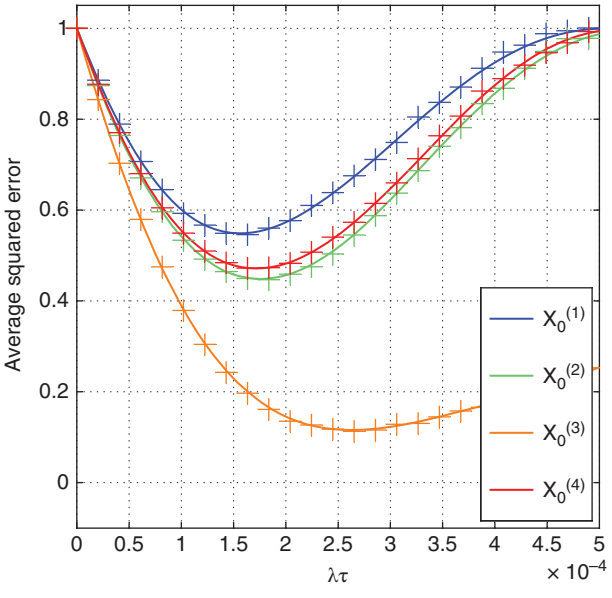


Figure 7.1 Comparison of the risk estimate using Monte Carlo (solid line) and SURE (cross) versus $\lambda \times \tau$ for $\mathbf{X}_0 \in \mathbb{R}^{200\times500}$. SNR = 0.5

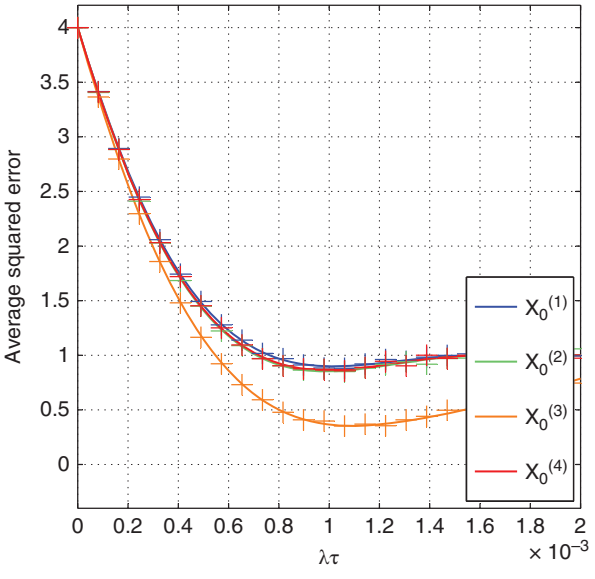


Figure 7.2 The same as Figure 7.1 except SNR = 1

7.3.3 Extensions to Spectral Functions

Consider estimators given by spectral functions. These act on the singular values and take the form

$$f(\mathbf{Y}) = \sum_{i=1}^{\min(m,n)} f_i\left(\sigma_i\right)\mathbf{u}_i\mathbf{v}_i^T = \mathbf{U}f\left(\mathbf{\Sigma}\right)\mathbf{V}^H, \quad \text{for all } \mathbf{Y} \in \mathbb{R}^{m\times n} \tag{7.26}$$

where $f(\mathbf{Y}) = \mathbf{U}f(\mathbf{\Sigma})\mathbf{V}^H$ is any SVD (SVST is in this class). These functions admit a SURE formula, given by

$$\text{SURE}\left(f\right)(\mathbf{Y}) = -mn\tau^2 + \|f(\mathbf{Y}) - \mathbf{Y}\|_F^2 + 2\tau^2 \,\text{div}\left(f(\mathbf{Y})\right)$$

and that under mild assumptions there exists a closed form for their divergence:

$$\text{div}\left(f(\mathbf{Y})\right) = \sum_{i=1}^{\min(m,n)} \left(f_i^{'}\left(\sigma_i\right) + |m-n| \frac{f_i^{'}\left(\sigma_i\right)}{\sigma_i} \right) + 2 \sum_{i\neq j, i,j=1}^{\min(m,n)} \frac{\sigma_i f_i\left(\sigma_i\right)}{\sigma_i^2 - \sigma_j^2} \tag{7.27}$$

This is of interest because such estimators arise naturally in regularized regression problems. For instance, let $J : \mathbb{R}^{m\times n} \mapsto \mathbb{R}$ be a lower semicontinuous, proper convex function of the form

$$J(\mathbf{X}) = \sum_{i=1}^{\min(m,n)} J_i\left(\sigma_i(\mathbf{X})\right)$$

Then, for $\lambda > 0$ the estimator

$$f_\lambda(\mathbf{Y}) = \underset{\mathbf{X}\in\mathbb{R}^{m\times n}}{\arg\min} \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\|_F^2 + \lambda J(\mathbf{X}) \tag{7.28}$$

is spectral. We say a function is spectral if it depends on eigenvalues only.

An approach to recursively estimate the quadratic risk for matrix recovery problems regularized with spectral functions [386]. A class of matrix valued functions defined by singular values of nonsymmetric matrices is shown to have many properties analogous to matrix valued functions defined by eigenvalues of symmetric matrices. The strong semismoothness of singular values of a nonsymmetric matrix is discussed and used to analyze the quadratic convergence of Newton's method for solving the inverse singular value problem.

Oymak and Hassibi [387] provided a sharp analysis of the minimax denoising problem and established a relation between the minimax MSE and phase transitions for arbitrary convex and continuous functions. Phase transitions deals with recovering a signal $\mathbf{x}_0$ from compressed linear observations $\mathbf{A}\mathbf{x}_0$ by minimizing a certain convex function $f(\cdot)$. On the other hand, minimax denoising is the problem of optimally estimating a signal $\mathbf{x}_0$ from noisy observations

$$\mathbf{y} = \mathbf{x}_0 + \mathbf{z}$$

using the regularization

$$\text{minimize} \quad \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|_2^2 + \lambda f(\mathbf{x})$$

where $||\cdot||_2$ is the Euclidean norm of a vector. In general, these problems are more meaningful and useful when the signal $\mathbf{x}_0$ has a certain structure and the convex function $f(\cdot)$ is chosen to exploit this structure. Examples of $f(\cdot)$ include, ℓ_1 and $\ell_1 - \ell_2$ norms for sparse and block sparse vectors, and nuclear norm $||\cdot||_*$ for low-rank matrices.

When the noise vector $\mathbf{z}$ is i.i.d. Gaussian, it is shown in [388] that the normalized estimation error (MSE) of the optimally tuned problem coincides with the compressed sensing phase transitions: the number $\Delta_f\left(\mathbf{x}_0\right)$ is such that one needs $m > \Delta_f\left(\mathbf{x}_0\right)$ compressed observations $\mathbf{A}\mathbf{x}_0 \in \mathbb{R}^m$ to recover the signal $\mathbf{x}_0$ by solving

$$\begin{aligned} &\text{minimize} \quad f(\mathbf{x}) \\ &\text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{x}_0 \end{aligned}$$

$\Delta_f(\mathbf{x}_0)$ can be obtained as an explicit formula based on the subdifferential of $f(\cdot)$ at $\mathbf{x}_0$.

Following [389], we suppose we observe a single noisy matrix $\mathbf{Y}$, generated by adding noise $\mathbf{Z}$ to an unknown matrix $\mathbf{X}_0$, so that

$$\mathbf{Y} = \mathbf{X}_0 + \mathbf{Z}$$

where $\mathbf{Z}$ is a noise matrix. Our goal is to recover the matrix $\mathbf{X}_0$ with some bound on the mean-squared error (MSE). This is hopeless if $\mathbf{X}_0$ is a completely general matrix and the noise $\mathbf{Z}$ is arbitrary; but if $\mathbf{X}_0$ happens to be of relatively low rank and the noise matrix $\mathbf{Z}$ is i.i.d standard Gaussian, we can indeed guarantee quantitatively accurate recovery. Donoho and Gavish (2013) provided explicit formulas for the best possible guarantees obtainable by a popular, computationally practical procedure.

Let $\mathbf{Y}, \mathbf{X}_0$, and $\mathbf{Z}$ be $m \times m$ matrices and suppose that $\mathbf{Z}$ has i.i.d entries, $Z_{ij} \sim \mathcal{N}(0,1)$. Consider the following nuclear-norm penalization problem:

$$\hat{\mathbf{X}}_\lambda = \underset{\mathbf{X} \in \mathcal{M}_{m\times n}}{\arg\min} \frac{1}{2} \|\mathbf{Y} - \mathbf{X}\|_F^2 + \lambda \|\mathbf{X}\|_* \tag{7.29}$$

where $\|\mathbf{X}\|_*$ denotes the sum of singular values of $\mathbf{X} \in \mathbb{C}^{m\times n}$, also known as the nuclear norm, and $\lambda > 0$ is a penalty factor. A solution to (7.29) is efficiently computable by modern convex optimization software CVX [378]; it shrinks away from $\mathbf{Y}$ in the direction of a smaller nuclear norm.

Measure performance (risk) is by the use of mean-squared error (MSE). When the unknown $\mathbf{X}_0$ is of known rank r and belongs to a matrix class $\mathcal{X}_{m\times n} \subset \mathcal{M}_{m\times n}$, the minimax MSE of nuclear-norm penalization is

$$\mathcal{M}_{m\times n}(r\,|\mathcal{X}) = \inf_{\lambda>0} \sup_{\substack{\mathbf{X}_0 \in \mathcal{X}_{m,n} \\ \text{rank}(\mathbf{X}_0) \leqslant r}} \frac{1}{mn} \mathbb{E}_{\mathbf{X}_0} \left\| \hat{\mathbf{X}}_\lambda (\mathbf{X}_0 + \mathbf{Z}) - \mathbf{X}_0 \right\|_F^2 \tag{7.30}$$

namely the worst case risk of $\hat{\mathbf{X}}_{\lambda_\star}$ where $\lambda_\star$ is the threshold for which this worst case risk is the smallest possible. In a very clear sense $\mathcal{M}_{m\times n}(r\,|\mathcal{X})$ gives the best possible guarantee for the MSE of nuclear-norm penalization (7.29), based solely on the rank and problem size, and not on other properties of the matrix $\mathbf{X}_0$.

7.3.4 Regularized Principal Component Analysis

Principal component analysis (PCA) is a well established dimensionality reduction method commonly used to denoise and visualize data. Regularized PCA [390] is relevant to the SURE method. The SURE method relies on a soft thresholding strategy:

$$\text{SVST}_\lambda(\mathbf{Y}) = \sum_{i=1}^{\min(m,n)} (\sigma_i - \lambda)_+ \mathbf{u}_i \mathbf{v}_i^T$$

The threshold parameter λ is automatically selected by minimizing Stein's unbiased risk estimate (SURE). As a tuning parameter, the SURE method does not require the number of underlying dimensions of the signal but it does require estimation of the noise variance σ^2 to determine λ.

When data can be seen as a true signal corrupted by error, PCA does not provide the best recovery of the underlying signal. Shrinking the singular values improves the

estimation of the underlying structure especially when data are noisy. Soft thresholding is one of the most popular strategies and involves linearly shrinking the singular values. The regularized PCA suggested in [390] applies a nonlinear transformation of the singular values associated with a hard thresholding rule. The regularized term is analytically derived from the MSE using asymptotic results from nonlinear regression models or using Bayesian considerations.

7.4 Asymptotic Framework for Matrix Reconstruction

The purpose of this section is to introduce the method of using asymptotic limits of large random matrices for matrix estimation. Random matrix theory underlies this method. It is remarkable that the algorithm developed using this method is asymptotically optimal and often optimal in practice. This method appears promising for future smart-grid power systems and big data.

By studying the asymptotic framework, we focus on the deterministic aspects of the problem, which is analogous to the method of studying the expectation of random variables (scalar, vector, matrix or tensor).

7.4.1 Matrix Estimation with Loss Functions

We address the problem of recovering a low rank signal matrix whose entries are observed in the presence of additive Gaussian noise [391]. Our goal is to recover an unknown $m \times n$ matrix $\mathbf{X}_0$ of low rank that is observed in the presence of i.i.d. Gaussian noise as matrix Y:

$$\mathbf{Y} = \mathbf{X}_0 + \frac{\sigma}{\sqrt{n}}\mathbf{Z}, \qquad \text{where } Z_{ij} \sim \text{i.i.d. } \mathcal{N}(0,1)$$

The factor $1/\sqrt{n}$ ensures that the signal and noise are comparable, and is employed for the asymptotic study of matrix reconstruction. We can consider the variance of the noise σ^2 to be known, and assume that it is equal to one. We can also obtain an estimator for σ, which we use in the proposed reconstruction method. In this case we have

$$\mathbf{Y} = \mathbf{X}_0 + \frac{1}{\sqrt{n}}\mathbf{Z}, \qquad \text{where } Z_{ij} \sim \text{i.i.d. } \mathcal{N}(0,1) \tag{7.31}$$

Formally, a matrix recovery scheme is a map $g : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ from the space of $m \times n$ matrices to itself. Given a recovery scheme $g(\cdot)$ and an observed matrix $\mathbf{Y}$ from the model (7.31), we regard $\hat{\mathbf{X}}_0 = g(\mathbf{Y})$ as an estimate of $\mathbf{X}_0$, and measure the performance of the estimate $\hat{\mathbf{X}}_0$ by

$$\text{Loss}\left(\mathbf{X}_0, \hat{\mathbf{X}}_0\right) = \left\|\mathbf{X}_0 - \hat{\mathbf{X}}_0\right\|_F^2 \tag{7.32}$$

where $||\cdot||_F$ denotes the Frobenius norm. The Frobenius norm of an $m \times n$ matrix $\mathbf{A} = \{a_{ij}\}$ is given by

$$\|\mathbf{A}\|_F^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2$$

Note that the vector space $\mathbb{R}^{m\times n}$ is equipped with the inner product $\langle \mathbf{A}, \mathbf{B}\rangle = \mathrm{Tr}\left(\mathbf{A}^T\mathbf{B}\right)$, then $\langle \mathbf{A}, \mathbf{A}\rangle = \|\mathbf{A}\|_F^2$.

A natural starting point for reconstruction of the target matrix A in (7.31) is the singular value decomposition (SVD) of the observed matrix $\mathbf{Y}$. Recall that the singular value decomposition of an $m \times n$ matrix $\mathbf{Y}$ is given by the factorization

$$\mathbf{Y} = \mathbf{U}\mathbf{D}\mathbf{V}^T = \sum_{i=1}^{\min(m,n)} d_i \mathbf{u}_i \mathbf{v}_i^T$$

Here $\mathbf{U}$ is an $m \times n$ orthogonal matrix whose columns are the left singular vectors $\mathbf{u}_i$, $\mathbf{V}$ is an $n \times n$ orthogonal matrix whose columns are the right singular vectors $\mathbf{v}_i$, and $\mathbf{D}$ is an $m \times n$ matrix with singular values $d_i = D_{ii} \geqslant 0$ on the diagonal and all other entries equal zero.

Many matrix reconstruction schemes act by shrinking the singular values of the observed matrix towards zero. Shrinkage is typically accomplished by hard or soft thresholding. Hard thresholding schemes set every singular value of $\mathbf{Y}$ less than a given positive threshold λ equal to zero, leaving other singular values unchanged. The family of hard thresholding schemes is defined by

$$g_\lambda^H(\mathbf{Y}) = \sum_{i=1}^{\min(m,n)} d_i I\left(d_i \geqslant \lambda\right) \mathbf{u}_i \mathbf{v}_i^T, \qquad \lambda > 0$$

where $I(\cdot)$ is the indicator function. Soft thresholding schemes subtract a given positive number ν from each singular value, setting values less than ν equal to zero. The family of soft thresholding schemes is defined by

$$g_\lambda^S(\mathbf{Y}) = \sum_{i=1}^{\min(m,n)} \left(d_i - \nu\right)_+ \mathbf{u}_i \mathbf{v}_i^T, \qquad \nu > 0$$

Hard and soft thresholding schemes can be defined equivalently in the penalized forms

$$g_\lambda^H(\mathbf{Y}) = \arg\min_{\mathbf{A}} \left\{\|\mathbf{Y} - \mathbf{A}\|_F^2 + \lambda^2 \,\mathrm{rank}(\mathbf{A})\right\}$$
$$g_\lambda^S(\mathbf{Y}) = \arg\min_{\mathbf{A}} \left\{\|\mathbf{Y} - \mathbf{A}\|_F^2 + 2\nu\|\mathbf{A}\|_*\right\}$$

where $||\mathbf{A}||_F$ denotes the nuclear norm of $\mathbf{A}$, which is equal to the sum of its singular values.

We now deal with orthogonally invariant reconstruction methods. The additive model (7.31) and Frobenius loss (7.32) have several elementary invariance properties, that lead naturally to the consideration of reconstruction methods with analogous forms of invariance. Recall that a square matrix $\mathbf{U}$ is said to be orthogonal if $\mathbf{U}\mathbf{U}^T = \mathbf{U}^T\mathbf{U} = \mathbf{I}$, or equivalently, if the rows (or columns) of $\mathbf{U}$ are orthonormal. If we multiply each side of (7.31) from the left side and right side, respectively, by orthogonal matrices $\mathbf{U}$ and $\mathbf{V}^T$ of appropriate dimensions, we obtain

$$\mathbf{U}\mathbf{Y}\mathbf{V}^T = \mathbf{U}\mathbf{X}_0\mathbf{V}^T + \frac{1}{\sqrt{n}}\mathbf{U}\mathbf{Z}\mathbf{V}^T \tag{7.33}$$

(7.33) is a reconstruction problem of the form (7.31) with signal $\mathbf{U}\mathbf{X}_0\mathbf{V}^T$ and observed matrix $\mathbf{U}\mathbf{Y}\mathbf{V}^T$. If $\hat{\mathbf{X}}_0$ is an estimate of $\mathbf{X}_0$ in model (7.31), then $\mathbf{U}\hat{\mathbf{X}}_0\mathbf{V}^T$ is an estimate of $\mathbf{U}\mathbf{X}_0\mathbf{V}^T$ in model (7.33) with the same loss. We have

$$\begin{aligned}\text{Loss}\left(\mathbf{U}\mathbf{X}_0\mathbf{V}^T, \mathbf{U}\hat{\mathbf{X}}_0\mathbf{V}^T\right) &= \left\|\mathbf{U}\left(\mathbf{X}_0 - \hat{\mathbf{X}}_0\right)\mathbf{V}^T\right\|_F^2 \\ &= \left\|\mathbf{X}_0 - \hat{\mathbf{X}}_0\right\|_F^2 = \text{Loss}\left(\mathbf{X}_0, \hat{\mathbf{X}}_0\right)\end{aligned}$$

A random $m \times n$ matrix Z has an orthogonally invariant distribution if for any orthogonal matrices $\mathbf{U}$ and $\mathbf{V}$ of appropriate size the distribution of $\mathbf{U}\mathbf{Z}\mathbf{V}^T$ is the same as the distribution of $\mathbf{Z}$.

It is natural to consider reconstruction schemes whose action does not change under orthogonal transformations of the reconstruction problem. A reconstruction scheme $g(\mathbf{Y})$ is orthogonally invariant if for any $m \times n$ matrix $\mathbf{Y}$, and any orthogonal matrices $\mathbf{U}$ and $\mathbf{V}$ of appropriate size

$$g\left(\mathbf{U}\mathbf{Z}\mathbf{V}^T\right) = \mathbf{U}g(\mathbf{Y})\mathbf{V}^T$$

Theorem 7.4.1 Let $\mathbf{Y} = \mathbf{A} + \mathbf{W}$, where $\mathbf{A}$ is a random target matrix. Assume that $\mathbf{A}$ and $\mathbf{W}$ are independent and have orthogonally invariant distributions. Then, for every reconstruction scheme $g(\cdot)$, there is an orthogonally invariant reconstruction scheme $\tilde{g}(\cdot)$, whose expected loss is the same, or smaller, than that of $g(\cdot)$.

The next proposition follows from our ability to diagonalize the signal matrix $\mathbf{A}$ in the reconstruction problem.

Proposition 7.4.2 Let $\mathbf{Y} = \mathbf{A} + \frac{1}{\sqrt{n}}\mathbf{W}$, where $\mathbf{W}$ has an orthogonally invariant distribution. If $g(\cdot)$ is an orthogonally invariant reconstruction scheme, then for any fixed signal matrix $\mathbf{A}$, the distribution of $\text{Loss}\left(\mathbf{A}, g(\mathbf{Y})\right)$, and in particular $\mathbb{E}\,\text{Loss}\left(\mathbf{A}, g(\mathbf{Y})\right)$, depends only on the singular values of $\mathbf{A}$.

Proof. Let $\mathbf{U}\mathbf{D}_A\mathbf{V}^T$ be the SVD of $\mathbf{A}$. Then $\mathbf{D}_A = \mathbf{U}^T\mathbf{A}\mathbf{V}$, and as the Frobenius norm is invariant under left and right orthogonal multiplications

$$\begin{aligned}\text{Loss}\left(\mathbf{A}, g(\mathbf{Y})\right) &= \|g(\mathbf{Y}) - \mathbf{A}\|_F^2 = \left\|\mathbf{U}^T\left(g(\mathbf{Y}) - \mathbf{A}\right)\mathbf{V}\right\|_F^2 \\ &= \left\|\mathbf{U}^T g(\mathbf{Y})\mathbf{V} - \mathbf{U}^T\mathbf{A}\mathbf{V}\right\|_F^2 = \left\|g\left(\mathbf{U}^T\mathbf{Y}\mathbf{V}\right) - \mathbf{D}_A\right\|_F^2 \\ &= \left\|g\left(\mathbf{D}_A + \tfrac{1}{\sqrt{n}}\mathbf{U}^T\mathbf{W}\mathbf{V}\right) - \mathbf{D}_A\right\|_F^2\end{aligned}$$

The result now follows from the fact that $\mathbf{U}^T\mathbf{W}\mathbf{V}$ has the same distribution as $\mathbf{W}$. □

We now address the implications of our ability to diagonalize the observed matrix $\mathbf{Y}$. Let $g(\cdot)$ be an orthogonally invariant reconstruction method, and let $\mathbf{U}\mathbf{D}\mathbf{V}^T$ be the singular value decomposition of $\mathbf{Y}$. It follows from the orthogonal invariance of $g(\cdot)$ that

$$g(\mathbf{Y}) = g\left(\mathbf{U}\mathbf{Z}\mathbf{V}^T\right) = \mathbf{U}g\left(\mathbf{D}\right)\mathbf{V}^T = \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}\mathbf{u}_i\mathbf{v}_j^T \tag{7.34}$$

where c_{ij} depend only on the singular values of $\mathbf{Y}$. In particular, any orthogonally invariant $g(\cdot)$ reconstruction method is completely determined by how it acts on diagonal matrices. The following theorem allows us to substantially refine the representation (7.34).

Theorem 7.4.3 Let $g(\cdot)$ be an orthogonally invariant reconstruction scheme. Then $g(\mathbf{Y})$ is diagonal whenever $\mathbf{Y}$ is diagonal.

As an immediate corollary of Theorem 7.4.3 and Equation (7.34) we obtain a compact, and useful, representation of any orthogonally invariant reconstruction scheme $g(\cdot)$.

Corollary 7.4.4 Let $g(\mathbf{Y})$ be an orthogonally invariant reconstruction method. If the observed matrix $\mathbf{Y}$ has singular value decomposition $\mathbf{Y} = \sum_{i=1}^{\min(m,n)} d_i \mathbf{u}_i \mathbf{v}_i^T$ then the reconstructed matrix has the form

$$\hat{\mathbf{A}} = g(\mathbf{Y}) = \sum_{i=1}^{\min(m,n)} c_i \mathbf{u}_i \mathbf{v}_i^T \tag{7.35}$$

where the coefficients c_i depend only on the singular values of $\mathbf{Y}$.

The converse of Corollary 7.4.4 is true under a mild additional condition. Let $g(\mathbf{Y})$ be a reconstruction scheme such that $g(\mathbf{Y}) = \sum_i c_i \mathbf{u}_i \mathbf{v}_i^T$, where $c_i = c_i\left(d_1, \ldots, d_{\min(m,n)}\right)$ are fixed functions of the singular values of $\mathbf{Y}$. If the functions $\{c_i(\cdot)\}$ are such that $c_i(d) = c_j(d)$ whenever $d_i = d_j$, then $g(\cdot)$ is orthogonally invariant. This follows from the uniqueness of the singular value decomposition.

Now we connect asymptotic matrix reconstruction with random matrix theory. Random matrix theory deals roughly with the *spectral* properties (i.e. eigenvalues) of random matrices, and is an obvious starting point for an analysis of matrix reconstruction. Using recent results on spiked population models, following [391], we establish asymptotic connections between the singular values and vectors of the signal matrix $\mathbf{A}$ and those of the observed matrix $\mathbf{Y}$. These asymptotic connections provide us with finite-sample estimates that can be applied in a nonasymptotic setting to matrices of small or moderate dimensions.

7.4.2 Connection with Large Random Matrices

The proposed reconstruction method is derived from an asymptotic version of the matrix reconstruction problem (7.31). For $n \geq 1$ let integers $m = m(n)$ be defined in such a way that

$$\frac{m}{n} \to c > 0 \text{ as } n \to \infty \tag{7.36}$$

For each n, let $\mathbf{Y}, A$ and $\mathbf{W}$ be $m \times n$ matrices such that

$$\mathbf{Y} = \mathbf{A} + \frac{1}{\sqrt{n}}\mathbf{W} \tag{7.37}$$

where the entries of $\mathbf{W}$ are independent $\mathcal{N}(0,1)$ random variables. We assume that the signal matrix $\mathbf{A}$ has fixed rank $r \geq 0$ and fixed non-zero singular values $\sigma_1(\mathbf{A}), \ldots, \sigma_r(\mathbf{A})$

that are independent of n. The constant c represents the limiting aspect ratio of the observed matrices $\mathbf{Y}$. The scale factor $\frac{1}{\sqrt{n}}$ ensures that the singular values of the signal matrix are comparable to those of the noise.

Proposition 7.4.5 Under the asymptotic reconstruction model with $\mathbf{A} = 0$ the empirical distribution of the singular values $\sigma_1(\mathbf{Y}) \geqslant \cdots \geqslant \sigma_{\min(m,n)}(\mathbf{A})$ converges weakly to a (non-random) limiting distribution with density

$$f_{\mathbf{Y}}(t) = \frac{1}{\pi \min(1,c)} \frac{1}{\sqrt{t}} \sqrt{(a-t^2)(t^2-b)}, \quad t \in \left[\sqrt{a}, \sqrt{b}\right] \tag{7.38}$$

where $a = (1-\sqrt{c})^2$ and $b = (1+\sqrt{c})^2$. Moreover, $\sigma_1(\mathbf{Y}) \xrightarrow{\mathbb{P}} 1+\sqrt{c}$ and $\sigma_{\min(m,n)}(\mathbf{Y}) \xrightarrow{\mathbb{P}} 1-\sqrt{c}$ as n tends to infinity.

The existence and form of the density $f_{\mathbf{Y}}(\cdot)$ are a consequence of the classical Marchenko–Pastur theorem [172, 173]. If $c = 1$, the density function $f_{\mathbf{Y}}(\cdot)$ simplifies to the quarter-circle law $f_{\mathbf{Y}}(t) = \pi^{-1}\sqrt{4-t^2}$, for $t \in [0,2]$.

The next two results concern the limiting eigenvalues and eigenvectors of $\mathbf{Y}$ when $\mathbf{A}$ is nonzero. Proposition 7.4.6 relates the limiting eigenvalues of $\mathbf{Y}$ to the (fixed) eigenvalues of $\mathbf{A}$, while Proposition 7.4.7 relates the limiting singular vectors of $\mathbf{Y}$ to the singular vectors of $\mathbf{A}$.

Proposition 7.4.6 ([392]) If $\mathbf{Y}$ follow the asymptotic matrix reconstruction model (7.37) with signal singular values $\sigma_1(\mathbf{A}) \geqslant \cdots \geqslant \sigma_r(\mathbf{A}) > 0$. For $1 \leq i \leq r$, as n tends to infinity

$$\sigma_i(\mathbf{Y}) \xrightarrow{\mathbb{P}} \begin{cases} \left(1+\sigma_i^2(\mathbf{A})+c+\frac{c}{\sigma_i^2(\mathbf{A})}\right)^{1/2} & \text{if } \sigma_i(\mathbf{A}) > c^{1/4} \\ 1+\sqrt{c} & \text{if } 0 < \sigma_i(A) \leqslant c^{1/4}. \end{cases}$$

The remaining singular values $\sigma_{r+1}(\mathbf{Y}), \ldots, \sigma_{\min(m,n)}(\mathbf{Y})$ of $\mathbf{Y}$ are associated with the zero singular values of $\mathbf{A}$: their empirical distribution converges weakly to the limiting distribution in Proposition 7.4.5.

Proposition 7.4.7 ([393–395]) Let $\mathbf{Y}$ follow the asymptotic matrix reconstruction model (7.37) with distinct signal singular values $\sigma_1(\mathbf{A}) > \cdots > \sigma_r(\mathbf{A}) > 0$. Fix i such that $\sigma_i(\mathbf{A}) > c^{1/4}$. Then, as n tends to infinity

$$\langle \mathbf{u}_i(\mathbf{Y}), \mathbf{u}_i(\mathbf{A}) \rangle^2 \xrightarrow{\mathbb{P}} \left(1-\frac{c}{\sigma_i^4(\mathbf{A})}\right) \Big/ \left(1+\frac{c}{\sigma_i^4(\mathbf{A})}\right)$$

and

$$\langle \mathbf{v}_i(\mathbf{Y}), \mathbf{v}_i(\mathbf{A}) \rangle^2 \xrightarrow{\mathbb{P}} \left(1-\frac{c}{\sigma_i^4(\mathbf{A})}\right) \Big/ \left(1+\frac{c}{\sigma_i^4(\mathbf{A})}\right)$$

Moreover, if $j = 1, \ldots, r$ not equal to i then $\left\langle \mathbf{u}_i(\mathbf{Y}), \mathbf{u}_j(\mathbf{A}) \right\rangle \xrightarrow{\mathbb{P}} 0$ and $\left\langle \mathbf{v}_i(\mathbf{Y}), \mathbf{v}_j(\mathbf{A}) \right\rangle \xrightarrow{\mathbb{P}} 0$ as n tends to infinity.

The limits established in Proposition 7.4.6 indicate a phase transition. If the singular value $\sigma_i(\mathbf{A})$ is less than or equal to $c^{1/4}$ asymptotically, the singular value $\sigma_i(\mathbf{Y})$ lies within the support of the Marchenko–Pastur distribution and is not distinguishable from the noise singular values. On the other hand, if the singular value $\sigma_i(\mathbf{A})$ exceeds $c^{1/4}$, then, asymptotically, $\sigma_i(\mathbf{Y})$lies outside the support of the Marchenko–Pastur distribution, and the corresponding left and right singular vectors of $\mathbf{Y}$ are associated with those of $\mathbf{A}$ (Proposition 7.4.7).

7.4.3 Asymptotic Matrix Reconstruction

Assume for the moment that the variance σ^2 of the noise is known, and equal to one. Let $\mathbf{Y}$ be an observed $m \times n$ matrix generated from the additive model $\mathbf{Y} = \mathbf{A} + \frac{1}{\sqrt{n}}\mathbf{W}$, and let

$$\mathbf{Y} = \sum_{i=1}^{\min(m,n)} \sigma_i(\mathbf{Y})\mathbf{u}_i(\mathbf{Y})\mathbf{v}_i^T(\mathbf{Y})$$

be the SVD of $\mathbf{Y}$. We seek an estimate $\hat{\mathbf{A}}$ of the signal matrix $\mathbf{A}$ having the form

$$\hat{\mathbf{A}} = \sum_{i=1}^{\min(m,n)} c_i\mathbf{u}_i(\mathbf{Y})\mathbf{v}_i^T(\mathbf{Y})$$

with each coefficient c_i depending only on the singular values $\sigma_1(\mathbf{Y}), \ldots, \sigma_{\min(m,n)}(\mathbf{Y})$ of $\mathbf{Y}$. We derive $\hat{\mathbf{A}}$ from the limiting relations in Propositions 7.4.6 and 7.4.7. By way of approximation, we treat these relations as exact in the nonasymptotic setting under study, using the symbol $\overset{l}{=}$, $\overset{l}{\leqslant}$, and $\overset{l}{>}$ to denote limiting equality and inequality relations.

Suppose initially that the singular values and vectors of the signal matrix A are known. In this case we wish to obtain coefficients $\{c_i\}$ minimizing

$$\begin{aligned}\text{Loss}\left(\mathbf{A}, \hat{\mathbf{A}}\right) &= \left\|\hat{\mathbf{A}} - \mathbf{A}\right\|_F^2 \\ &= \left\|\sum_{i=1}^{\min(m,n)} c_i\mathbf{u}_i(\mathbf{Y})\mathbf{v}_i^T(\mathbf{Y}) - \sum_{i=1}^{r} \sigma_i(\mathbf{A})\mathbf{u}_i(\mathbf{A})\mathbf{v}_i^T(\mathbf{A})\right\|_F^2\end{aligned}$$

Proposition 7.4.6 shows that asymptotically the information about the singular values of $\mathbf{A}$ that are smaller than $c^{1/4}$ is not recoverable from the singular values of $\mathbf{Y}$. As a result we can restrict the first sum to the first $r_0 = \#\left\{i : \sigma_i(\mathbf{A}) > c^{1/4}\right\}$ terms

$$\text{Loss}\left(\mathbf{A}, \hat{\mathbf{A}}\right) = \left\|\sum_{i=1}^{r_0} c_i\mathbf{u}_i(\mathbf{Y})\mathbf{v}_i^T(\mathbf{Y}) - \sum_{i=1}^{r} \sigma_i(\mathbf{A})\mathbf{u}_i(\mathbf{A})\mathbf{v}_i^T(\mathbf{A})\right\|_F^2$$

Proposition 7.4.7 ensures that the left singular vectors $\mathbf{u}_i(\mathbf{A})$ and $\mathbf{u}_j(\mathbf{A})$ are asymptotically orthogonal for $i = 1, \ldots, r$ not equal to $j = 1, \ldots, r_0$ and therefore

$$\text{Loss}\left(\mathbf{A}, \hat{\mathbf{A}}\right) \overset{l}{=} \sum_{i=1}^{r_0} \left\|c_i\mathbf{u}_i(\mathbf{Y})\mathbf{v}_i^T(\mathbf{Y}) - \sigma_i(\mathbf{A})\mathbf{u}_i(\mathbf{A})\mathbf{v}_i^T(\mathbf{A})\right\|_F^2 + \sum_{i=r_0+1}^{r} \sigma_i^2(\mathbf{A})$$

Fix $1 \leq i \leq r_0$. Expanding the i-th term in the above sum gives

$$\begin{aligned}&\left\|\sigma_i(\mathbf{A})\mathbf{u}_i(\mathbf{A})\mathbf{v}_i^T(\mathbf{A}) - c_i\mathbf{u}_i(\mathbf{Y})\mathbf{v}_i^T(\mathbf{Y})\right\|_F^2 \\ &= c_i^2\left\|\mathbf{u}_i(\mathbf{Y})\mathbf{v}_i^T(\mathbf{Y})\right\|_F^2 + \sigma_i^2(\mathbf{A})\left\|\mathbf{u}_i(\mathbf{A})\mathbf{v}_i^T(\mathbf{A})\right\|_F^2\end{aligned}$$

$$
\begin{aligned}
&- 2c_i\sigma_i(\mathbf{A})\left\langle \mathbf{u}_i(\mathbf{A})\mathbf{v}_i^T(\mathbf{A}), \mathbf{u}_i(\mathbf{Y})\mathbf{v}_i^T(\mathbf{Y})\right\rangle \\
&= \sigma_i^2(\mathbf{A}) + c_i^2 - 2c_i\sigma_i(\mathbf{A})\left\langle \mathbf{u}_i(\mathbf{A}), \mathbf{u}_i(\mathbf{Y})\right\rangle \left\langle \mathbf{v}_i(\mathbf{A}), \mathbf{v}_i(\mathbf{Y})\right\rangle
\end{aligned}
$$

Differentiating the last expression with respect to c_i yields the optimal value

$$
c_i^\star = \sigma_i(\mathbf{A})\left\langle \mathbf{u}_i(\mathbf{A}), \mathbf{u}_i(\mathbf{Y})\right\rangle \left\langle \mathbf{v}_i(\mathbf{A}), \mathbf{v}_i(\mathbf{Y})\right\rangle \tag{7.39}
$$

In order to estimate the coefficient $c_i^\star$, we consider separately singular values of $\mathbf{Y}$ that are at most, or greater than $1+\sqrt{c}$, where $c = m/n$ is the aspect ratio of $\mathbf{Y}$. By Proposition 7.4.6, the asymptotic relation $\sigma_i(\mathbf{Y}) \overset{l}{\leqslant} 1+\sqrt{c}$ implies $\sigma_i(\mathbf{A}) \leqslant c^{1/4}$, and in this case the i-th singular value of $\mathbf{A}$ is not recoverable from $\mathbf{Y}$. Thus if $\sigma_i(\mathbf{Y}) \leqslant 1+\sqrt{c}$, we set the corresponding coefficient $c_i^\star = 0$.

On the other hand, the asymptotic relation $\sigma_i(\mathbf{Y}) \overset{l}{>} 1+\sqrt{c}$ implies $\sigma_i(\mathbf{A}) > c^{1/4}$, and that each of the inner products in (7.39) is asymptotically positive. The displayed equations in Propositions 7.4.6 and 7.4.7 can then be used to obtain estimates of each term in (7.39) based only on the (observed) singular values of $\mathbf{Y}$ and its aspect ratio c. These equations yield the following relations:

$$
\begin{aligned}
\hat{\sigma}_i^2(\mathbf{A}) &= \frac{1}{2}\left[\sigma_i^2(\mathbf{Y}) - (1+c) + \sqrt{\left[\sigma_i^2(\mathbf{Y}) - (1+c)\right]^2 - 4c}\right] \text{ estimates } \sigma_i^2(\mathbf{A}) \\
\hat{\theta}_i^2 &= \left(1 - \frac{c}{\hat{\sigma}_i^4(\mathbf{A})}\right) \Big/ \left(1 + \frac{c}{\hat{\sigma}_i^4(\mathbf{A})}\right) \text{ estimates } \left\langle \mathbf{u}_i(\mathbf{A}), \mathbf{u}_i(\mathbf{Y})\right\rangle^2 \\
\hat{\phi}_i^2 &= \left(1 - \frac{c}{\hat{\sigma}_i^4(\mathbf{A})}\right) \Big/ \left(1 + \frac{c}{\hat{\sigma}_i^4(\mathbf{A})}\right) \text{ estimates } \left\langle \mathbf{v}_i(\mathbf{A}), \mathbf{v}_i(\mathbf{Y})\right\rangle^2
\end{aligned}
$$

With these estimates in hand, the matrix reconstruction scheme is defined via the equation

$$
G_o^{RMT}(\mathbf{Y}) = \sum_{\sigma_i(\mathbf{A}) > 1+\sqrt{c}} \hat{\sigma}_i(\mathbf{A})\hat{\theta}_i\hat{\phi}_i\mathbf{u}_i(\mathbf{Y})\mathbf{v}_i^T(\mathbf{Y}) \tag{7.40}
$$

where $\hat{\sigma}_i(\mathbf{A})$, $\hat{\theta}_i$ and $\hat{\phi}_i$ are the positive square roots of the estimates defined above.

The RMT method shares features with both hard and soft thresholding. It sets to zero singular values of $\mathbf{Y}$ smaller than the threshold $(1+\sqrt{c})$, and it shrinks the remaining singular values towards zero. However, unlike soft thresholding, the amount of shrinkage depends on the singular values, the larger singular values are shrunk less than the smaller ones. Unlike hard and soft thresholding schemes, the proposed RMT method has no tuning parameters. The only unknown, the noise variance, is estimated within the procedure.

In the general version of the matrix reconstruction problem, the variance σ^2 of the noise is not known. In this case, given an estimate $\hat{\sigma}^2$ of σ^2, such as that described below, we may define

$$
G^{RMT}(\mathbf{Y}) = \hat{\sigma} G_o^{RMT}\left(\frac{\mathbf{Y}}{\hat{\sigma}}\right) \tag{7.41}
$$

7.4.4 Estimation of the Noise Variance

Now let us show how an estimate $\hat{\sigma}^2$ of σ^2 is obtained. Let $\mathbf{Y}$ be derived from the asymptotic reconstruction model $\mathbf{Y} = \mathbf{A} + \sigma\frac{1}{\sqrt{n}}\mathbf{W}$, with σ unknown.

A function $f(\cdot) : \mathbb{R}^{m\times n} \to \mathbb{R}$ is orthogonally invariant if for any $m \times n$ matrix $\mathbf{Y}$ and any orthogonal matrices $\mathbf{U}$ and $\mathbf{V}$ of appropriate sizes,

$$f(\mathbf{Y}) = f\left(\mathbf{U}\mathbf{Y}\mathbf{V}^T\right) \tag{7.42}$$

Proposition 7.4.8 A function $f(\cdot) : \mathbb{R}^{m\times n} \to \mathbb{R}$ is orthogonally invariant if and only if $f(\mathbf{Y})$ depends **only** on the singular values of $\mathbf{Y}$.

Proposition 7.4.9 Let $f(\cdot) : \mathbb{R}^{m\times n} \to \mathbb{R}$. Then there is an orthogonally invariant function $\tilde{f}(\cdot)$ with the following property. Let $\mathbf{A}$ and $\mathbf{W}$ be independent $m \times n$ matrices with orthogonally invariant distributions, and let $\mathbf{Y} = \mathbf{A} + \sigma\frac{1}{\sqrt{n}}\mathbf{W}$ for some σ. Then $\tilde{f}(\mathbf{Y})$ has the same expected value as $f(\mathbf{Y})$ and a smaller or equal variance.

Let F be the cumulative distribution function of the density (7.38). For each $\sigma > 0$, let $\hat{S}_\sigma$ be the set of singular values $\sigma_i(\mathbf{Y})$ that fall in the interval $\left[\sigma\left|1-\sqrt{c}\right|, \sigma\left(1+\sqrt{c}\right)\right]$, and let $\hat{F}_\sigma$ be the empirical cumulative distribution function of $\hat{S}_\sigma$. Then

$$K(\sigma) = \sup_s \left|F(s/\sigma) - \hat{F}_\sigma(s)\right|$$

is the Kolmogorov–Smirnov distance between the empirical and theoretical singular value distribution functions [396], and we define

$$\hat{\sigma}_i(\mathbf{Y}) = \arg\min_{\sigma>0} K(\sigma) \tag{7.43}$$

to be the value of σ minimizing $K(\sigma)$. A routine argument shows that the estimator $\hat{\sigma}$ is scale invariant in the sense that $\hat{\sigma}(\alpha\mathbf{Y}) = \alpha\hat{\sigma}(\mathbf{Y})$ for each $\alpha > 0$.

By considering the jump points of the empirical cumulative distribution function $\hat{F}_\sigma(s)$, the supremum in $K(\sigma)$ simplifies to

$$K(\sigma) = \max_{s_i\in\hat{S}_\sigma}\left|F\left(s_i/\sigma\right) - \frac{i-1/2}{\left|\hat{S}_\sigma\right|}\right| + \frac{1}{2\left|\hat{S}_\sigma\right|} \tag{7.44}$$

where $\{s_i\}$ are the ordered elements of $\hat{S}_\sigma$. The objective function $K(\sigma)$ is discontinuous at points where the $\hat{S}_\sigma$ changes, so one minimizes it over a fine grid of points σ in the range where $\left|\hat{S}_\sigma\right| > \frac{1}{2}\min(m,n)$ and $\sigma\left(1+\sqrt{c}\right) < 2\sigma_1(\mathbf{Y})$.

The closed form of the cumulative distribution function $F(\cdot)$ is calculated as the integral of $f_{\mathbf{W}/\sqrt{n}}(t)$, which is defined in (7.38). For $c = 1 (a = 0, b = 4)$ it is a common integral

$$F(t) = \int_{\sqrt{a}}^{t} f(x)\,dx = \frac{1}{\pi}\int_0^t \sqrt{b-x^2}dx = \frac{1}{2\pi}\left(t\sqrt{4-t^2} + 4\arcsin\frac{t}{2}\right)$$

For $c \neq 1$ the calculations are more complicated. First we perform the change of variables, which yields

$$\begin{aligned} F(t) &= \int_{\sqrt{a}}^{t} f(x)dx = C\int_{\sqrt{a}}^{t}\frac{1}{x^2}\sqrt{(b-x^2)(x^2-a)}dx \\ &= \int_a^{t^2}\frac{1}{y}\sqrt{(b-y)(y-a)}dy \end{aligned}$$

where $C = 1/(2\pi\min(c,1))$. See [391] to find the closed form of $F(t)$.

7.4.5 Optimal Hard Threshold for Matrix Denoising

Suppose we are interested in an $m \times n$ matrix $\mathbf{X}$, which is thought to be either exactly or approximately of low rank, but we only observe a single noisy $m \times n$ matrix $\mathbf{Y}$, obeying the additive model $\mathbf{Y} = \mathbf{A} + \sigma\mathbf{Z}$; The noise matrix $\mathbf{Z}$ has independent, identically distributed entries with zero mean and unit variance. We wish to recover the matrix $\mathbf{X}$ with some bound on the mean squared error (MSE). The default estimation technique for our task is *Truncated* SVD (TSVD) [397]: Write

$$\mathbf{Y} = \sum_{i=1}^{m} y_i \mathbf{u}_i \mathbf{v}_i^T \tag{7.45}$$

for the singular value decomposition of the data matrix $\mathbf{Y}$, where $\mathbf{u}_i \in \mathbb{R}^m$, and $\mathbf{v}_i \in \mathbb{R}^n$, $i = 1, \ldots, m$ are the left and right singular vectors of $\mathbf{Y}$ corresponding to the singular value y_i. The TSVD estimator is

$$\hat{\mathbf{X}}_r = \sum_{i=1}^{r} y_i \mathbf{u}_i \mathbf{v}_i^T$$

where $r = \text{rank}(\mathbf{X})$, assumed known, and $y_1 \geqslant \cdots \geqslant y_m$. Being the best approximation of rank r to the data in the least squares sense and therefore the maximum likelihood estimator when $\mathbf{Z}$ has Gaussian entries, the TSVD is arguably as ubiquitous in science and engineering as linear regression.

When the true rank r of the signal $\mathbf{X}$ is unknown, one might try to form an estimate of the rank $\hat{r}$ and then apply the TSVD $\hat{\mathbf{X}}_r$. It is instructive to think about rank estimation (using any method), followed by TSVD, simply as hard thresholding of the data singular values, where only components $y_i \mathbf{u}_i \mathbf{v}_i^T$ for which y_i passes a specified threshold, are included in $\hat{\mathbf{X}}_r$. Let

$$\eta_H\left(y; \tau\right) = y\mathbf{1}_{\{y > \tau\}}$$

denote the hard thresholding nonlinearity, and consider singular value hard thresholding (SVHT)

$$\hat{\mathbf{X}}_\tau = \sum_{i=1}^{m} \eta_H\left(y_i; \tau\right) \mathbf{u}_i \mathbf{v}_i^T \tag{7.46}$$

In words, $\hat{\mathbf{X}}_\tau$ sets to 0 any data singular value below τ.

Let us measure the denoising performance of a denoiser $\hat{\mathbf{X}}$ at a signal matrix using mean square error

$$\left\| \hat{\mathbf{X}}(\mathbf{Y}) - \mathbf{Y} \right\|_F^2 = \sum_{i,j} \left(\hat{X}(\mathbf{Y})_{ij} - X_{ij} \right)^2$$

The TSVD is an optimal rank-r approximation of the data matrix $\mathbf{Y}$. But this does not necessarily mean that it is a good, or even reasonable, estimator for the signal matrix $\mathbf{X}$.

Following [391], we adopt an asymptotic framework where the matrix grows while keeping the nonzero singular values of $\mathbf{X}$ fixed, and the signal-to-noise ratio of those singular values stays constant with increasing n.

In this asymptotic framework, for a low-rank $n \times n$ matrix observed in white (not necessarily Gaussian) noise of level σ

$$\tau_\star = \frac{4}{\sqrt{3}}\sqrt{n}\sigma \approx 2.309\sqrt{n}\sigma$$

is the optimal location for the hard thresholding of singular values. For a nonsquare $m \times n$ matrix with $m \neq n$, the optimal location is

$$\tau_\star = \gamma_\star(c) \cdot \sqrt{n}\sigma \tag{7.47}$$

where $c = m/n$. The value $\gamma_\star$ is the optimal hard threshold coefficient for known σ. It is given by

$$\gamma_\star(c) = \sqrt{2(c+1) + \frac{8c}{(c+1) + \sqrt{c^2 + 14c + 1}}} \tag{7.48}$$

Many authors have considered matrix denoising by applying the soft thresholding nonlinearity

$$\eta_S(y; s) = \left(|y| - s\right)_+ \cdot \text{sign}(y)$$

instead of hard thresholding, to the data singular values. The denoiser

$$\hat{\mathbf{X}}_{soft} = \sum_{i=1}^{n} \eta_S(y; s)\, \mathbf{u}_i \mathbf{v}_i^T$$

is known as singular value soft thresholding (SVST) or SVT.

Consider an optimal singular value shrinker. In the asymptotic framework we are using, Shabalin and Nobel [391]—see also Section 7.4.3—have derived an optimal singular value shrinker $\hat{\mathbf{X}}_{opt}$. Calibrated for the model $\mathbf{Y} = \mathbf{X} + \frac{1}{\sqrt{n}}\mathbf{Z}$, in the square setting $m = n$, this shrinker takes the form

$$\hat{\mathbf{X}}_{opt} : \sum_{i=1}^{n} y_i \mathbf{u}_i \mathbf{v}_i^T \mapsto \sum_{i=1}^{n} \eta_{opt}(y_i)\, \mathbf{u}_i \mathbf{v}_i^T$$

where

$$\eta_{opt}(t) = \sqrt{(t^2 - 4)_+}$$

In our asymptotic framework, this rule dominates, in asymptotic mean square error (AMSE), any other estimator based on singular value shrinkage, at any configuration of the singular values of the low-rank signal.

In our asymptotic framework, this thresholding rule adapts to unknown ranks and, if needed, to unknown noise levels, in an optimal manner: it is *always* better than hard thresholding at any other value, no matter what the matrix is that we are trying to recover, and is *always* better than ideal truncated SVD (TSVD), which truncates at the true rank of the low-rank matrix we are trying to recover.

Hard thresholding at the recommended value to recover an $n \times n$ matrix of rank r guarantees an AMSE at most $3nr\sigma^2$. In comparison, the guarantee provided by TSVD is $5nr\sigma^2$, the guarantee provided by optimally tuned singular value soft thresholding is $6nr\sigma^2$, and the best guarantee achievable by any shrinkage of the data singular values is $2nr\sigma^2$. Our recommended hard threshold value also offers, among hard thresholds, the

best possible AMSE guarantees for recovering matrices with bounded nuclear norms. Empirical evidence shows that these AMSE properties of the $4/\sqrt{3}$ thresholding rule remain valid even for relatively small n, and that performance improvement over TSVD and other popular shrinkage rules is often substantial, turning it into the practical hard threshold of choice.

Example 7.4.10 (optimal singular value hard thresholding in practice) For a low-rank $n \times n$ matrix observed in white (not necessarily Gaussian) noise of unknown level, one can use the data to obtain an approximation of the optimal location $\tau_\star$. Define

$$\tau_\star \approx 2.858 \cdot y_{median}$$

where y_{median} is the median singular value of the data matrix $\mathbf{Y}$. The notation $\tau_\star$ is meant to emphasize that this is not a fixed threshold chosen a priori, but rather a data-dependent threshold. For a nonsquare $m \times n$ matrix with $m \neq n$, the approximate optimal location when σ is unknown is

$$\hat{\tau}_\star = \omega(c) \cdot y_{median} \tag{7.49}$$

where $\omega(c)$ is approximated by

$$\omega(c) \approx 0.56c^3 - 0.95c^2 + 1.82c + 1.43 \tag{7.50}$$

The accurate computation can be done using the MATLAB script provided in the original paper [398].

The optimal SVHT for unknown noise level, $\hat{\mathbf{X}}_{\hat{\tau}_\star}$, is very simple to implement and does not require any tuning parameters. The denoised matrix $\hat{\mathbf{X}}_{\hat{\tau}_\star}(\mathbf{Y})$ can be computed using just a few code lines in a high-level scripting language. For example, in Matlab:

```
beta = size(Y,1) / size(Y,2)
omega = 0.56*beta^3 - 0.95*beta^2 + 1.82*beta + 1.43
[U D V] = svd(Y)
y = diag(Y)
y( y < (omega * median(y) ) ) = 0
Xhat = U * diag(y) * V'
```

In our asymptotic framework, $\tau_\star$ and $\hat{\tau}_\star$ enjoy exactly the same optimality properties. This means that $\hat{\mathbf{X}}_{\hat{\tau}_\star}$ adapts to unknown low rank *and* to unknown noise level. Empirical evidence suggest that their performance for finite n is similar. □

7.5 Optimum Shrinkage

Consider $m \times n$ signal-plus-noise data or measurement matrix

$$\mathbf{Y} = \mathbf{A} + \mathbf{X} \tag{7.51}$$

with

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i(\mathbf{A}) \mathbf{u}_i(\mathbf{A}) \mathbf{v}_i^T(\mathbf{A}) \tag{7.52}$$

where $\mathbf{u}_i(\mathbf{A})$ and $\mathbf{v}_i(\mathbf{A})$ are left and right "signal" singular vectors associated with singular values $\sigma_i(\mathbf{A})$ of the signal matrix $\mathbf{A}$ with rank r, and $\mathbf{X}$ is the noise-only matrix of random (not necessarily i.i.d.) noises.

As pointed out above, when the rank r is known, the truncated singular value decomposition (SVD) plays a prominent role in a widely used "optimal" solution to a problem that is addressed by the famous Eckart–Young–Mirsky theorems [397,399,400]. We call this solution truncated SVD

$$\hat{\mathbf{A}}^{\text{TSVD}} = \underset{\text{rank}(\mathbf{A})=r}{\arg\min} \|\mathbf{Y} - \mathbf{A}\|_F \tag{7.53}$$

where

$$\mathbf{Y} = \sum_{i=1}^{\min(m,n)} \sigma_i(\mathbf{Y})\mathbf{u}_i(\mathbf{Y})\mathbf{v}_i^T(\mathbf{Y})$$

is the SVD of the noisy matrix $\mathbf{Y}$. This is also the maximum likelikhood (ML), rank r estimate when $\mathbf{X}$ is assumed to be a matrix with i.i.d. Gaussian entries because the negative log-likelihood function is precisely the right-hand side of (7.53). Its use is also justified in the small m, large n (or vice versa) regime, whenever local asymptotic normality [401] has "kicked in."

The solution to the optimization problem

$$\mathbf{w}^{\text{TSVD}} := \underset{\|\mathbf{w}\|_{\ell_0}=r}{\arg\min} \left\| \sum_{i=1}^{r} \sigma_i(\mathbf{A})\mathbf{u}_i(\mathbf{A})\mathbf{v}_i^T(\mathbf{A}) - \sum_{i=1}^{\min(m,n)} w_i\mathbf{u}_i(\mathbf{Y})\mathbf{v}_i^T(\mathbf{Y}) \right\|_F \tag{7.54}$$

is given by

$$w_i^{\text{TSVD}} = \sigma_i(\mathbf{Y}), i = 1, \ldots, r$$

This yields the rank r signal matrix estimate

$$\hat{\mathbf{A}}^{\text{TSVD}} = \sum_{i=1}^{r} w_i^{\text{TSVD}}\mathbf{u}_i(\mathbf{Y})\mathbf{v}_i^H(\mathbf{Y})$$

which, by the Eckart–Young–Mirsky theorem, is also the solution to the representation problem in (7.54).

Now suppose no structure is assumed in the low-rank matrix. Let $\|\mathbf{w}\|_{\ell_0} = \left|\left\{\#i : w_i \neq 0\right\}\right|$ so that $\|\mathbf{w}\|_{\ell_0} = r$ denotes a vector $\mathbf{w}$ with r nonzero entries.

$$\mathbf{w}^{opt} := \underset{\|\mathbf{w}\|_{\ell_0}=r}{\arg\min} \left\| \sum_{i=1}^{r} \sigma_i(\mathbf{A})\mathbf{u}_i(\mathbf{A})\mathbf{v}_i^T(\mathbf{A}) - \sum_{i=1}^{\min(m,n)} w_i\mathbf{u}_i(\mathbf{Y})\mathbf{v}_i^T(\mathbf{Y}) \right\|_F \tag{7.55}$$

The analysis shows that $\mathbf{w}^{opt}$ takes the form of a shrinkage-and-thresholding operator (on the singular values of $\mathbf{Y}$) that is completely characterized by the limiting singular value distribution of the noise-only matrix. The resulting shrinkage function is nonconvex with

$$w_i^{opt} \approx \sigma_i(\mathbf{Y})\left[1 - O\left(1/\sigma_i^2(\mathbf{Y})\right)\right]$$

for large $\sigma_i(\mathbf{Y})$ and

$$w_i^{opt} \to 0, \text{ for } \sigma_i(\mathbf{Y}) \leqslant b + o(1)$$

where b is a critical threshold that depends on the limiting noise-only singular value distribution.

7.6 A Shrinkage Approach to Large-Scale Covariance Matrix Estimation

Many applied problems require a covariance matrix estimator that is not only invertible but is also well-conditioned (that is, inverting it does not amplify estimation error). For large dimensional covariance matrices, the usual estimator, the sample covariance matrix, is typically not well conditioned and may not even be invertible. Ledoit and Wolf (2004) [402] introduced an estimator that is both well conditioned and more accurate than the sample covariance matrix asymptotically. This estimator is distribution free and has a simple explicit formula that is easy to compute and interpret. It is the asymptotically optimal convex linear combination of the sample covariance matrix with the identity matrix. Optimality is meant with respect to a quadratic loss function, asymptotically as both the number of observations n and the number of variables p go to infinity together (called general asymptotics). The only constraint is the ratio p/n must remain bounded. Extensive Monte Carlo simulations indicate that 20 observations and variables are enough for the asymptotic approximations to typically hold well in a finite sample.

The empirical (sample) covariance matrix $\mathbf{S}$ can not anymore be considered a good approximation of the true covariance matrix $\mathbf{\Sigma}$ (this is true also for moderately sized data with $n \sim p$).

The easiest way to explain what we do is first to analyze in detail the finite sample case. Let $\mathbf{X}$ denote a $p \times n$ matrix of n independent and identically distributed (i.i.d) observations on a system of p random variables with mean zero and covariance matrix $\mathbf{\Sigma}$. We consider the Frobenius norm

$$\|\mathbf{A}\|_F^2 = \frac{1}{p} \operatorname{Tr}\left(\mathbf{A}\mathbf{A}^H\right) = \frac{1}{p} \sum_{i=1}^{n} \sum_{j=1}^{n} \left|a_{ij}\right|^2 = \frac{1}{p} \sum_{i=1}^{n} \lambda_i^2(\mathbf{A}) \tag{7.56}$$

Dividing by p is not standard, but it does not matter here because p remain finite. The advantages of this convention are that the norm of the identity matrix is simply one and that it will be consistent. The norm of the p_n-dimensional matrix $\mathbf{A}$ is:$\|\mathbf{A}\|_F^2 = \operatorname{Tr}\left(\mathbf{A}\mathbf{A}^H\right)/p$.

Our goal is to find the linear combination $\mathbf{\Sigma}^* = \rho_1 \mathbf{I} + \rho_2 \mathbf{S}$ of the identity matrix $\mathbf{I}$ and the sample covariance matrix $\mathbf{S} = \mathbf{X}\mathbf{X}^H/n$ whose expected quadratic loss $\mathbb{E}\left[\left\|\mathbf{\Sigma}^* - \mathbf{\Sigma}\right\|_F^2\right]$ is minimum.

The squared Frobenius norm $\|\cdot\|_F^2$ is a quadratic form whose associated matrix inner product is:

$$\langle \mathbf{A}, \mathbf{B} \rangle = \operatorname{Tr}\left(\mathbf{A}\mathbf{B}^H\right)/p$$

In analogy with the inner product of two vectors, the matrix inner product can be viewed as the similarity between two matrices. Four scalars play a central role in the analysis:

$$\mu = \langle \mathbf{A}, \mathbf{I} \rangle, \alpha^2 = \|\mathbf{\Sigma} - \mu\mathbf{I}\|_F^2, \beta^2 = \|\mathbf{\Sigma} - \mathbf{S}\|_F^2, \text{and} \quad \delta^2 = \|\mathbf{S} - \mu\mathbf{I}\|_F^2$$

We do not need to assume that the entries (random variables) in $\mathbf{X}$ follow a specific distribution but we do need to assume that they have finite fourth moments, so that α^2 and β^2 are finite. It follows that

$$\begin{aligned}\mathbb{E}\left\|\mathbf{S}-\mu\mathbf{I}\right\|_F^2 &= \mathbb{E}\left\|\mathbf{S}-\boldsymbol{\Sigma}+\boldsymbol{\Sigma}-\mu\mathbf{I}\right\|_F^2 \\ &= \mathbb{E}\left\|\mathbf{S}-\boldsymbol{\Sigma}\right\|_F^2+\mathbb{E}\left\|\boldsymbol{\Sigma}-\mu\mathbf{I}\right\|_F^2+2\mathbb{E}\left\langle\mathbf{S}-\boldsymbol{\Sigma},\boldsymbol{\Sigma}-\mu\mathbf{I}\right\rangle \\ &= \mathbb{E}\left\|\mathbf{S}-\boldsymbol{\Sigma}\right\|_F^2+\mathbb{E}\left\|\boldsymbol{\Sigma}-\mu\mathbf{I}\right\|_F^2+2\left\langle\mathbb{E}\left[\mathbf{S}-\boldsymbol{\Sigma}\right],\boldsymbol{\Sigma}-\mu\mathbf{I}\right\rangle\end{aligned}$$

Note that $\mathbb{E}[\mathbf{S}]=\boldsymbol{\Sigma}$; therefore, the third term on the right-hand side of line 3, is equal to zero. Thus we have proven

$$\left\|\boldsymbol{\Sigma}-\mu\mathbf{I}\right\|_F^2+\left\|\boldsymbol{\Sigma}-\mathbf{S}\right\|_F^2=\left\|\mathbf{S}-\mu\mathbf{I}\right\|_F^2$$

or

$$\alpha^2+\beta^2=\delta^2 \tag{7.57}$$

The optimal linear combination $\boldsymbol{\Sigma}^*=\rho_1\mathbf{I}+\rho_2\mathbf{S}$ of the identity matrix $\mathbf{I}$ and the sample covariance matrix $\mathbf{S}=\mathbf{X}\mathbf{X}^H/n$ is the standard solution to a simple quadratic programming problem under the linear equality constraint.

Theorem 7.6.1 Consider the optimization problem:

$$\begin{aligned}&\min_{\rho_1,\rho_2}\mathbb{E}\left[\left\|\boldsymbol{\Sigma}^*-\boldsymbol{\Sigma}\right\|_F^2\right] \\ &\text{subject to } \boldsymbol{\Sigma}^*=\rho_1\mathbf{I}+\rho_2\mathbf{S}\end{aligned} \tag{7.58}$$

where the coefficients ρ_1 and ρ_2 are nonrandom. Its solution verifies:

$$\boldsymbol{\Sigma}^*=\frac{\beta^2}{\delta^2}\mu\mathbf{I}+\frac{\alpha^2}{\delta^2}\mathbf{S} \tag{7.59}$$

$$\mathbb{E}\left[\left\|\boldsymbol{\Sigma}^*-\boldsymbol{\Sigma}\right\|_F^2\right]=\frac{\alpha^2\beta^2}{\delta^2} \tag{7.60}$$

See [402] for the proof of the above solution.

$\mu\mathbf{I}$ can be interpreted as a shrinkage target and the weight $\frac{\beta^2}{\delta^2}$ placed on $\mu\mathbf{I}$ as a shrinkage intensity. The percentage relative improvement in average loss (PRIAL) over the sample covariance matrix is equal to

$$\frac{\left[\left\|\mathbf{S}-\boldsymbol{\Sigma}\right\|_F^2\right]-\mathbb{E}\left[\left\|\boldsymbol{\Sigma}^*-\boldsymbol{\Sigma}\right\|_F^2\right]}{\mathbb{E}\left[\left\|\mathbf{S}-\boldsymbol{\Sigma}\right\|_F^2\right]}=\frac{\beta^2}{\delta^2} \tag{7.61}$$

the same as the shrinkage intensity. Therefore, everything is controlled by the ratio $\frac{\beta^2}{\delta^2}$, which is a properly normalized measure of the error of the sample covariance matrix $\mathbf{S}$. Intuitively, if $\mathbf{S}$ is relatively accurate, then you should not shrink it too much, and shrinking it will not help you much either; if $\mathbf{S}$ is relatively inaccurate, then you should shrink it a lot and you also stand to gain a lot from shrinking. It is well known that $\mathbf{S}$ is inaccurate when n and p are large and comparable, so shrinking will benefit us a lot.

The mathematics underlying Theorem 7.6.1 is so rich that we are able to provide four complementary interpretations of it. One is geometric and the others echo some of the most important ideas in finite sample multivariate statistics. First, we can see Theorem 7.6.1 as a projection theorem in Hilbert space. See Figure 7.3 for an illustration. The second way to interpret Theorem 7.6.1 is as a tradeoff between bias

and variance. See Figure 7.4 for an illustration. We seek to minimize the mean-squared error, which can be decomposed into variance and squared bias:

$$\mathbb{E}\left[\left\|\mathbf{\Sigma}^{*}-\mathbf{\Sigma}\right\|_{F}^{2}\right]=\mathbb{E}\left[\left\|\mathbf{\Sigma}^{*}-\mathbb{E}\left[\mathbf{\Sigma}^{*}\right]\right\|_{F}^{2}\right]+\mathbb{E}\left[\left\|\mathbb{E}\left[\mathbf{\Sigma}^{*}\right]-\mathbf{\Sigma}\right\|_{F}^{2}\right] \tag{7.62}$$

The mean-squared error of the shrinkage target $\mu\mathbf{I}$ is all bias and no variance, while for the sample covariance matrix S it is exactly the opposite: all variance and no bias. $\mathbf{\Sigma}^{*}$ represents the optimal tradeoff between error due to bias and error due to variance. The idea of a tradeoff between bias and variance was already central to the original James and Stein [372] shrinkage technique.

The third interpretation is Bayesian. $\mathbf{\Sigma}^{*}$ can be seen as the combination of two signals: prior information and sample information. Prior information states that the true covariance matrix $\mathbf{\Sigma}$ lies on the sphere centered around the shrinkage target $\mu\mathbf{I}$ with radius α : Sample information states that $\mathbf{\Sigma}$ lies on another sphere, centered around the sample covariance matrix S with radius β. Bringing together prior and sample information, $\mathbf{\Sigma}$ must lie on the intersection of the two spheres, which is a circle. At the

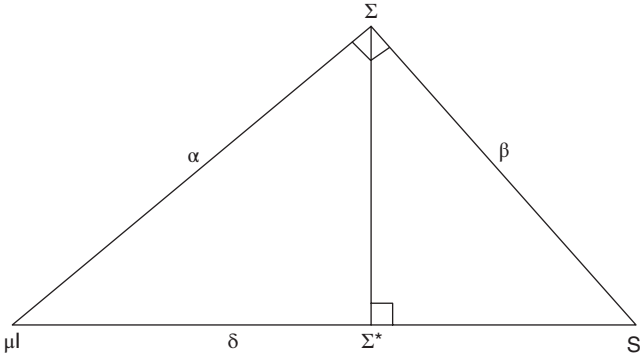
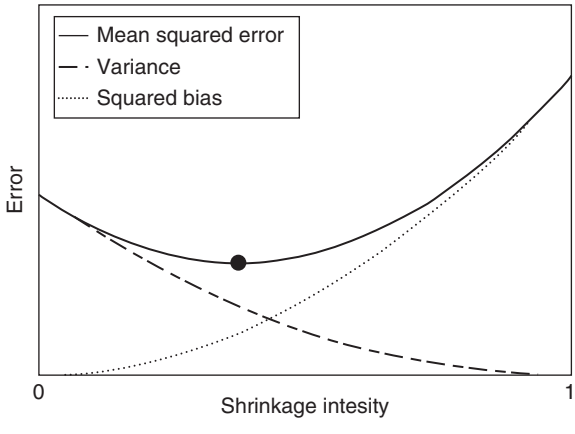


Figure 7.3 Theorem 7.6.1 interpreted as a projection in Hilbert space. Source: Reproduced from [402] with permission.

Figure 7.4 Theorem 7.6.1 interpreted as tradeoff between bias and variance: Shrinkage intensity 0 corresponds to the sample covariance matrix $\mathbf{S}$. Shrinkage intensity 1 corresponds to the shrinkage target $\mu\mathbf{I}$. Optimal shrinkage intensity (represented by •) corresponds to the minimum expected loss combination $\mathbf{\Sigma}^{*}$ Source: Reproduced from [402] with permission.



center of this circle stands $\mathbf{\Sigma}^*$. The relative importance given to prior vs. sample information in determining $\mathbf{\Sigma}^*$ depends on which one is more accurate. See Figure 7.5 for an illustration.

The fourth and last interpretation involves the cross-sectional dispersion of covariance matrix eigenvalues. Let $\lambda_1, \ldots, \lambda_p$ denote the eigenvalues of the true covariance matrix $\mathbf{\Sigma}$, and $\ell_1, \ldots, \ell_p$ those of the sample covariance matrix $\mathbf{S}$. We can exploit the Frobenius norm's elegant relationship to eigenvalues. Note that

$$\mu = \frac{1}{p}\sum_{i=1}^{p}\lambda_i = \mathbb{E}\left[\frac{1}{p}\sum_{i=1}^{p}\ell_i\right] \tag{7.63}$$

represents the grand mean of both true and sample eigenvalues. Then (7.57) can be rewritten as

$$\mathbb{E}\left[\frac{1}{p}\sum_{i=1}^{p}\left(\lambda_i-\mu\right)^2\right] = \frac{1}{p}\sum_{i=1}^{p}\left(\lambda_i-\mu\right)^2 + \mathbb{E}\left[\|\mathbf{S}-\mathbf{\Sigma}\|_F^2\right] \tag{7.64}$$

In words, sample eigenvalues are more dispersed around their grand mean than true ones, and the excess dispersion is equal to the error of the sample covariance matrix. Excess dispersion implies that the largest sample eigenvalues are biased upwards, and the smallest ones downwards. See Figure 7.6 for an illustration. Therefore, we can improve upon the sample covariance matrix by shrinking its eigenvalues towards their grand mean, as in

$$\forall i = 1, \ldots, p \quad \lambda_i^* = \frac{\beta^2}{\delta^2}\mu + \frac{\alpha^2}{\delta^2}\ell_i \tag{7.65}$$

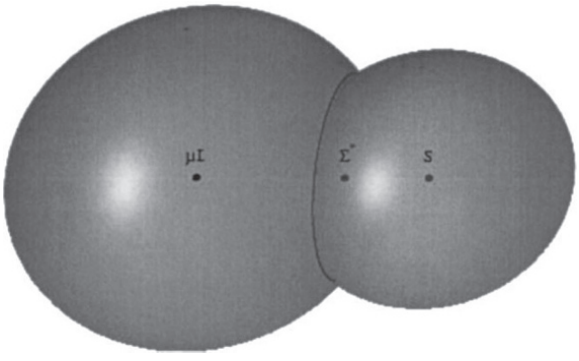


Figure 7.5 Bayesian interpretation. The left sphere has center $\mu\mathbf{I}$ and radius α and represents prior information. The right sphere has center $\mathbf{S}$ and radius β. The distance between sphere centers is δ and represents sample information. If all we knew was that the true covariance matrix $\mathbf{\Sigma}$ lies on the left sphere, our best guess would be its center: the shrinkage target $\mu\mathbf{I}$. If all we knew was that the true covariance matrix $\mathbf{\Sigma}$ lies on the right sphere, our best guess would be its center: the sample covariance matrix $\mathbf{S}$. Putting together both pieces of information, the true covariance matrix $\mathbf{\Sigma}$ must lie on the circle where the two spheres intersect; therefore, our best guess is its center: the optimal linear shrinkage $\mathbf{\Sigma}^*$. Source: Reproduced from [402] with permission.

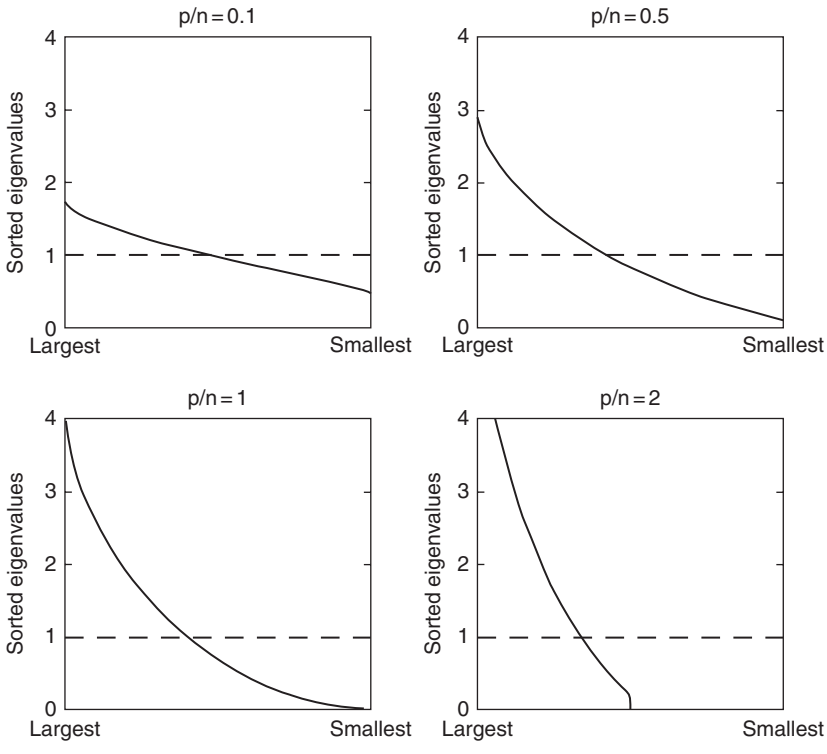


Figure 7.6 Sample versus true eigenvalues. The solid line represents the distribution of the eigenvalues of the sample covariance matrix. Eigenvalues are sorted from largest to smallest, then plotted against their rank. In this case, the true covariance matrix is the identity, that is, the true eigenvalues are all equal to one. The distribution of true eigenvalues is plotted as a dashed horizontal line at one. Distributions are obtained in the limit as the number of observations n and the number of variables p both go to infinity with the ratio p/n converging to a finite positive limit. The four plots correspond to different values of this limit. Source: Reproduced from [402] with permission.

Note that $\lambda_1^*, \ldots, \lambda_p^*$ defined in (7.65) are precisely the eigenvalues of $\mathbf{\Sigma}^*$. Surprisingly, their dispersion

$$\frac{1}{p}\sum_{i=1}^{p}\left(\lambda_i^* - \mu\right)^2 = \frac{\alpha^2}{\delta^2}$$

is even below the dispersion of true eigenvalues.

Example 7.6.2 (MATLAB Simulations) Computer code `covCor.m` in the Matlab programming language implementing this improved estimator [403] is freely downloadable from `http://www.ledoit.net`. (accessed August 6, 2016). □

Example 7.6.3 (Stein's phenomenon) A common key problem arises: how should one obtain an accurate and reliable estimate of the true covariance matrix $\mathbf{\Sigma}$ if presented with a data set that describes a large number of variables but only contains comparatively few samples ($n << p$)?

The simple solution is to rely either on the maximum likelihood estimate $\mathbf{S}^{ML}$ or on the related unbiased empirical covariance matrix $\mathbf{S} = \frac{n}{n-1}\mathbf{S}^{ML}$, with entries defined as

$$s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j)$$

where $\overline{x}_i = \frac{1}{n}\sum_{k=1}^{n} x_{ki}$ and x_{ki} is the k-th observation of the variable X_i.

Unfortunately both $\mathbf{S}$ and $\mathbf{S}^{ML}$ exhibit serious defects in the "small n, large p" data sets. Empirical covariance matrix can no longer be considered a good approximation of the true covariance matrix (this is also true for moderately sized data with $n \approx p$).

It has long been known that the two widely employed estimators of the covariance matrix, the unbiased $\mathbf{S}$ and the related maximum likelihood $\mathbf{S}^{ML}$ estimator, are both statistically inefficient. This is due to the so-called "Stein phenomenon" discovered by Stein (1956) [373] in the context of estimating the mean vector of a multivariate normal distribution. Stein demonstrated that in high-dimensional inference problems it is often possible to improve (sometimes dramatically!) upon the maximum likelihood estimator. This result is at first counterintuitive, as maximum likelihood can be proven to be *asymptotically* optimal, and as such it seems not unreasonable to expect that these favorable properties of maximum likelihood also extend to the case of finite data.

Further insight into the Stein effect is provided, [404], which points out that one needs to distinguish between two different aspects of maximum-likelihood inference. First, maximum likelihood as a means of summarizing the observed data and producing a maximum-likelihood summary (MLS). Second, maximum likelihood as a procedure to obtain a maximum-likelihood estimate (MLE). The conclusion is that maximum likelihood is unassailable as a data summarizer but that it has some clear defects as an estimating procedure.

This applies directly to the estimation of covariance matrices: $\mathbf{S}^{ML}$ constitutes the best estimator in terms of actual fit to the data but for medium to small data sizes it is far from being the optimal estimator for recovering the true covariance matrix $\mathbf{\Sigma}$. Fortunately, the Stein theorem also demonstrates that it is possible to construct a procedure for improved covariance estimation. □

Example 7.6.4 (squared Frobenius norm) The linear shrinkage approach suggests a weighted average

$$\mathbf{Q}^{\star} = \lambda\mathbf{T} + (1-\lambda)\mathbf{Q} \tag{7.66}$$

where $\lambda \in [0,1]$ denotes the shrinkage intensity. For $\lambda = 1$, the shrinkage estimate equals the shrinkage target $\mathbf{T}$, whereas for $\lambda = 0$ the unrestricted estimate $\mathbf{Q}$ is recovered. The key advantage of this construction is that it offers a systematic way to obtain a regularized estimate $\mathbf{Q}^{\star}$ that outperforms the individual estimators $\mathbf{Q}$ and $\mathbf{T}$ both in terms of accuracy and statistical efficiency.

In a matrix setting the equivalent to the squared error loss function is the squared Frobenius norm.

$$\begin{aligned} L(\lambda) &= \left\|\mathbf{S}^{\star} - \mathbf{\Sigma}\right\|_F^2 \\ &= \left\|\lambda\mathbf{T} + (1-\lambda)\mathbf{S} - \mathbf{\Sigma}\right\|_F^2 \\ &= \sum_{i=1}^{p}\sum_{j=1}^{p}(\lambda t_{ij} + (1-\lambda)s_{ij} - \sigma_{ij})^2 \end{aligned} \tag{7.67}$$

is a natural quadratic measure of distance between the true $\mathbf{\Sigma}$ and inferred covariance matrix ($\mathbf{S}^\star$). In this formula the unconstrained unbiased empirical covariance matrix $\mathbf{S}$ replaces the unconstrained estimate $\mathbf{Q}$ of (7.66).

It is advantageous to choose the parameter λ in a *data-driven* fashion by explicitly minimizing a risk function

$$R(\lambda) = \mathbb{E}[L(\lambda)]$$

It is less well known that the optimal regularization parameter λ may often also be determined *analytically*. Specifically, Ledoit and Wolf (2003) [405] derived a simple theorem for choosing λ that guarantees minimal MSE without the need of having to specify any underlying distributions, and without requiring computationally expensive procedures such as MCMC, the bootstrap, or crossvalidation.

The loss function is extremely intuitive: it is a quadratic measure of distance between the true and the estimated covariance matrices based on the Frobenius norm $\|\cdot\|_F$ defined in (7.56).

we have to choose the objective according to which the shrinkage intensity is "optimal." It follows from (7.67) that

$$\begin{aligned}
R(\lambda) &= \mathbb{E}\left[L(\lambda)\right] \\
&= \sum_{i=1}^{p}\sum_{j=1}^{p} \mathbb{E}\left(\lambda t_{ij} + (1-\lambda)s_{ij} - \sigma_{ij}\right)^2 \\
&= \sum_{i=1}^{p}\sum_{j=1}^{p} \operatorname{Var}\left(\lambda t_{ij} + (1-\lambda)s_{ij}\right) + \left[\mathbb{E}\left(\lambda t_{ij} + (1-\lambda)s_{ij} - \sigma_{ij}\right)\right]^2 \\
&= \sum_{i=1}^{p}\sum_{j=1}^{p} \lambda^2 \operatorname{Var}\left(t_{ij}\right) + (1-\lambda)^2 \operatorname{Var}\left(s_{ij}\right) \\
&\quad + 2\lambda(1-\lambda)\operatorname{Cov}\left(t_{ij}, s_{ij}\right) + \lambda^2\left(\phi_{ij} - \sigma_{ij}\right)^2
\end{aligned}$$

The goal now is to minimize the risk $R(\lambda)$ with respect to λ. Calculating the first two derivatives of $R(\lambda)$ yields, after some basic algebra

$$\begin{aligned}
R'(\lambda) &= 2\sum_{i=1}^{p}\sum_{j=1}^{p} \lambda \operatorname{Var}\left(t_{ij}\right) - (1-\lambda)\operatorname{Var}\left(s_{ij}\right) \\
&\quad + (1-2\lambda)\operatorname{Cov}\left(t_{ij}, s_{ij}\right) + \lambda\left(\phi_{ij} - \sigma_{ij}\right)^2 \\
R''(\lambda) &= 2\sum_{i=1}^{p}\sum_{j=1}^{p} \operatorname{Var}\left(t_{ij} - s_{ij}\right) + \left(\phi_{ij} - \sigma_{ij}\right)^2
\end{aligned}$$

where ϕ_{ij} is the (i,j) entry of some matrix $\mathbf{\Phi}$. See [406] for details. Setting $R'(\lambda)$ and solving for $\lambda^\star$ we get

$$\lambda^\star = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} \operatorname{Var}\left(s_{ij}\right) - \operatorname{Cov}\left(t_{ij}, s_{ij}\right)}{\sum_{i=1}^{p}\sum_{j=1}^{p} \operatorname{Var}\left(t_{ij} - s_{ij}\right) + \left(\phi_{ij} - \sigma_{ij}\right)^2} \tag{7.68}$$

$R''(\lambda)$ is positive everywhere, so this solution is verified as a minimum of our risk function. □

7.7 Eigenvectors of Large Sample Covariance Matrix Ensembles

Consider N independent samples $\mathbf{z}_1, \ldots, \mathbf{z}_N$, all of which are $n \times 1$ real or complex vectors $\mathbb{R}^n$ or $\mathbb{C}^n$. In this section, we are interested in the large-n-limiting spectral properties of the sample covariance matrix

$$\mathbf{S}_n = \frac{1}{N}\mathbf{Z}\mathbf{Z}^H, \quad \mathbf{Z} = \left[\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N\right]$$

when we assume that the sample size $N = N(n)$ satisfies $N/n \to \gamma$ as $n \to \infty$ for some $\gamma > 0$. This framework is known as large-dimensional asymptotics. Throughout the section, $\mathbf{1}$ denotes the indicator function of a set. We make the following assumptions: $\mathbf{z} = \mathbf{\Sigma}^{1/2}\mathbf{x}$ where

- (H_1) $\mathbf{X}$ is a $n \times N$ matrix of real or complex i.i.d random variables with zero mean, unit variance, and the 12th absolute central moment is bounded by a constant C, independent of n and N;
- (H_2) the true covariance matrix $\mathbf{\Sigma}$ is a n-dimensional random Hermitian positive definite matrix, independent of $\mathbf{X}$;
- (H_3) $n/N \to \gamma > 0$ as $n \to \infty$;
- (H_4) $\lambda_1, \ldots, \lambda_n$ is a system of eigenvalues of $\mathbf{\Sigma}$, and the empirical spectral distribution (e.s.d.) of the true covariance matrix given by

$$H_n(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[\lambda_i, \infty)}(\lambda) \tag{7.69}$$

converges almost surely to a nonrandom limit $H(\lambda)$ at every point of continuity of H. H defines a probability distribution function, whose support $\mathrm{supp}(H)$ is included in the compact interval $\left[h_1, h_2\right]$ with $0 < h_1 \leqslant h_2 < \infty$.

7.7.1 Stieltjes Transform

The aim of this section is to investigate the asymptotic properties of the eigenvectors of sample covariance matrices. In particular, we will quantify how the eigenvectors of the sample covariance matrix deviate from those of a true covariance matrix under large-dimensional asymptotics. This will enable us to characterize how the sample covariance matrix deviates **as a whole** (i.e. through its eigenvalues and its eigenvectors) from the true covariance matrix.

In this section, we denote $\left(\left(\lambda_1, \ldots, \lambda_n\right); \left(\mathbf{u}_1, \ldots, \mathbf{u}_n\right)\right)$ a system of eigenvalues and orthonormal eigenvectors of the sample covariance matrix

$$\mathbf{S} = \frac{1}{N}\mathbf{\Sigma}^{1/2}\mathbf{X}\mathbf{X}^H\mathbf{\Sigma}^{1/2}$$

Without loss of generality, we assume that the eigenvalues are sorted in decreasing order: $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$. We also denote by $\left(\mathbf{v}_1, \ldots, \mathbf{v}_n\right)$ a system of orthonormal eigenvectors of the true covariance matrix $\mathbf{\Sigma}$.

The Stieltjes transform of a nondecreasing function R is defined by

$$m_R(z) = \int_{-\infty}^{+\infty} (t - z)^{-1} dR(t)$$

for all z in $\mathbb{C}^+$, where $\mathbb{C}^+ = \{z \in \mathbb{C}, \mathrm{Im}(z) > 0\}$. The use of the Stieltjes transform is motivated by the following inversion formula: given any nondecreasing function R

$$R(b) - R(a) = \lim_{\eta \to \infty} \frac{1}{\pi} \int_a^b \mathrm{Im}\left[m_R(\xi + j\eta)\right] d\xi$$

which holds if R is continuous at a and b.

The asymptotic behavior of the eigenvalues is now quite well understood. The "global behavior" of the spectrum of $\mathbf{S}$ for instance is characterized through the empirical spectrum density, defined as:

$$F_n(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{[\lambda_i, +\infty)}(\lambda), \qquad \forall \lambda \in \mathbb{R}$$

The empirical spectrum density is usually described through its Stieltjes transform.

The first fundamental result concerning the asymptotic global behavior of the spectrum was obtained by Marchenko and Pastur in [219]. Their result was later made more precise, for example in [173, 174, 221, 407, 408], and recent results are surveyed , for example, in [39, 163, 176]. We quote the most recent version as given in [175].

Let

$$m_{F_n}(z) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\lambda_i - z} = \frac{1}{n} \mathrm{Tr}\left[(\mathbf{S} - z\mathbf{I})^{-1}\right]$$

where $\mathbf{I}$ denotes the $n \times n$ identity matrix.

Theorem 7.7.1 ([219]) Under Assumptions (H1)–(H4), for all $z \in \mathbb{C}^+$, $\lim_{n \to \infty} m_{F_n}(z) = m_F(z)$ almost certainly where

$$\forall z \in \mathbb{C}^+, \quad m_F(z) = \int_{-\infty}^{+\infty} \frac{1}{\left[1 - \gamma^{-1} - \gamma^{-1} z m_F(z)\right] t - z} dH(t) \tag{7.70}$$

Furthermore, the empirical spectrum density of the sample covariance matrix given by

$$F_n(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{[\lambda_i, +\infty)}(\lambda)$$

converges almost certainly to the nonrandom limit $F(\lambda)$ at all points of continuity of F.

In addition, [222] shows that the following limit exists:

$$\forall \lambda \in \mathbb{R} - \{0\}, \quad \lim_{z \in \mathbb{C}+ \to \lambda} m_F(z) \equiv \tilde{m}_F(\lambda) \tag{7.71}$$

They also prove that F has a continuous derivative, which is given by $F' = \frac{1}{\pi} \mathrm{Im}\left[\tilde{m}_F(\lambda)\right]$ on $(0, +\infty)$. More precisely, when $\gamma > 1$, $\lim_{z \in \mathbb{C}+ \to \lambda} m_F(z) \equiv \tilde{m}_F(\lambda)$ exists for all $\lambda \in \mathbb{R}$, F has a continuous derivative $F?$ on all of $\mathbb{R}$, and $F(\lambda)$ is identically equal to zero in a neighborhood of $\lambda = 0$. When $\gamma < 1$, the proportion of sample eigenvalues equal to zero is asymptotically $1 - \gamma$. In this case, it is convenient to introduce the empirical distribution function

$$G = \left(1 - \gamma^{-1}\right) \mathbf{1}_{[0,\infty)}(\lambda) + \gamma^{-1} F \tag{7.72}$$

which is the limit of empirical distribution function of the eigenvalues of the N-dimensional matrix $\frac{1}{N}\mathbf{X}^H\mathbf{\Sigma}\mathbf{X}$. Then

$$\lim_{z\in\mathbb{C}+\to\lambda} m_G(z) \equiv \tilde{m}_G(\lambda) \tag{7.73}$$

exists for all $\lambda \in \mathbb{R}$, G has a continuous derivative G' for all of $\mathbb{R}$, and $G(\lambda)$ is identically equal to 0 in a neighborhood of $\lambda = 0$. When Y is exactly equal to 1, further complications arise because the density of sample eigenvalues can be unbounded in a neighborhood of zero; for this reason we sometimes have to rule out the possibility that $\gamma = 1$.

The Marchenko–Pastur equation reveals much of the behavior of the eigenvalues of sample covariance matrices under large-dimensional asymptotics. It is also of interest to describe the asymptotic behavior of the eigenvectors. Such an issue is fundamental to statistics (for instance both eigenvalues and eigenvectors are of interest in principal components analysis), wireless communication [39, 136], and finance.

Much less is known about eigenvectors of sample covariance matrices. In the special case where $\mathbf{\Sigma} = \mathbf{I}$ and the X_{ij} are i.i.d. standard (real or complex) Gaussian random variables, it is well known that the matrix of sample eigenvectors is Haar distributed (on the orthogonal or unitary group). As far as we are aware these are the only ensembles for which the distribution of the eigenvectors is explicitly known. A random matrix $\mathbf{U}$ is said to be asymptotically Haar distributed if $\mathbf{Ux}$ is asymptotically uniformly distributed on the unit sphere for any nonrandom unit vector $\mathbf{x}$.

In the case where $\mathbf{\Sigma} \neq \mathbf{I}$, much less is known (see [409, 410]). One expects that the distribution of the eigenvectors is far from being rotation invariant. This is precisely the aspect with which this section is concerned.

Following [411], we present another approach to the study of eigenvectors of sample covariance matrices. Roughly speaking, we study "functionals" of the type

$$\begin{aligned} \forall z \in \mathbb{C}^+, \quad \Phi_n^g(z) &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\lambda_i - z} \sum_{j=1}^{n} \left|\mathbf{u}_j^H \mathbf{v}_j\right|^2 \times g\left(\tau_j\right) \\ &= \frac{1}{n}\operatorname{Tr}\left[(\mathbf{S} - z\mathbf{I})^{-1} g\left(\mathbf{\Sigma}\right)\right] \end{aligned} \tag{7.74}$$

where g is any real-valued univariate function satisfying suitable regularity conditions. By convention, $g\left(\mathbf{\Sigma}\right)$ is the matrix with the same eigenvectors as $\mathbf{\Sigma}$ and with eigenvalues $g\left(\tau_1\right), \ldots, g\left(\tau_n\right)$. These functionals are generalizations of the Stieltjes transform used in the Marchenko–Pastur equation. Indeed, one can rewrite the Stieltjes transform of the empirical spectrum density as:

$$\forall z \in \mathbb{C}^+, \quad m_{F_n}(z) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\lambda_i - z} \sum_{j=1}^{n} \left|\mathbf{u}_j^H \mathbf{v}_j\right|^2 \times 1 \tag{7.75}$$

The constant 1 that appears at the end of (7.75) can be interpreted as a weighting scheme placed on the true eigenvectors; specifically, it represents a constant weighting scheme. The generalization we introduce here shows how the sample covariance matrix relates to the true covariance matrix, or even any function of the true covariance matrix.

The main result of this section is the following theorem.

Theorem 7.7.2 ([411]) Assume that conditions (H1)–(H4) are satisfied. Let g be a (real valued) bounded function defined on $[h_1, h_2]$ with finitely many points of discontinuity. Then there exists a nonrandom function $\Phi^g(z)$ defined over $\mathbb{C}^+$ such that

$$\Phi_n^g(z) = \frac{1}{n} \operatorname{Tr}\left[(\mathbf{S} - z\mathbf{I})^{-1} g(\mathbf{\Sigma})\right]$$

converges almost surely to $\Phi^g(z)$ for all $z \in \mathbb{C}^+$. Furthermore, $\Phi^g(z)$ is given by

$$\forall z \in \mathbb{C}^+, \quad \Phi^g(z) = \int_{-\infty}^{+\infty} \frac{g(\tau)}{\left[1 - \gamma^{-1} - \gamma^{-1} z m_F(z)\right] \tau - z} dH(\tau) \tag{7.76}$$

One can first observe that as we move from a flat weighting scheme of $g \equiv 1$ to any arbitrary weighting scheme $g(\tau_i)$, the integration kernel

$$\frac{1}{\left[1 - \gamma^{-1} - \gamma^{-1} z m_F(z)\right] \tau - z}$$

remains unchanged. Therefore, (7.76) generalizes Marchenko and Pastur's foundational result.

The generalization of the Marchenko–Pastur equation allows the consideration of a few unsolved problems regarding the overall relationship between sample and true covariance matrices. The first of these questions is: how do the eigenvectors of the sample covariance matrix deviate from those of the true covariance matrix? By injecting functions g of the form $\mathbf{1}_{[\lambda_i, +\infty)}$ into (7.76), we quantify the asymptotic relationship between sample and true eigenvectors.

Another question is: how does the sample covariance matrix deviate from the true covariance matrix as a whole, and how can we modify it to bring it closer to the true covariance matrix? This is an important question in statistics, where a covariance matrix estimator that improves upon the sample covariance matrix is sought. By injecting the function $g(\tau) = \tau$ into (7.76), we find the optimal asymptotic bias correction for the eigenvalues of the sample covariance matrix. We also perform the same calculation for the *inverse* covariance matrix this time by taking $g(\tau) = 1/\tau$.

7.7.2 Sample versus Population Eigenvectors

Each sample eigenvector $\mathbf{u}_i$ lies in a space whose dimension is growing towards infinity. Thus the only way to know "where" it lies is to project it onto a known orthonormal basis that will serve as a *reference grid*. Given the nature of the problem, the most natural choice for this reference grid is the orthonormal basis formed by the true eigenvectors $(\mathbf{v}_1, \ldots, \mathbf{v}_n)$. Now we are dealing with the asymptotic behavior of

$$\mathbf{u}_i^H \mathbf{v}_j, \qquad \text{for all } i, j = 1, \ldots, n$$

that is, the projection of the sample eigenvectors onto the true eigenvectors. Yet as every eigenvector is identified up to multiplication by a scalar of modulus one, the argument (angle) of $\mathbf{u}_i^H \mathbf{v}_j$ is devoid of mathematical relevance. Therefore, we can focus instead on its square modulus $\left|\mathbf{u}_i^H \mathbf{v}_j\right|$ without loss of information. Another issue that arises is that of scaling. As

$$\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left|\mathbf{u}_i^H \mathbf{v}_j\right|^2 = \frac{1}{n^2} \sum_{i=1}^{n} \mathbf{u}_i^H \left(\sum_{j=1}^{n} \mathbf{v}_i^H \mathbf{v}_j \right) \mathbf{u}_i = \frac{1}{n^2} \sum_{i=1}^{n} \mathbf{u}_i^H \mathbf{u}_i = \frac{1}{n}$$

we study $n\left|\mathbf{u}_i^H\mathbf{v}_j\right|^2$ instead, so that its limit does not vanish under large-n asymptotics. We choose to use an indexation system where "eigenvalues serve as labels for eigenvectors," that is $\mathbf{u}_i$ is the eigenvector associated to the i-th largest eigenvalue λ_i.

All these considerations lead us to introduce the following key object:

$$\forall \lambda, \tau \in \mathbb{R}, \qquad \phi_n(\lambda, \tau) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left|\mathbf{u}_i^H \mathbf{v}_j\right|^2 \mathbf{1}_{[\lambda_i, \infty)}(\lambda) \times \mathbf{1}_{[\tau_j, \infty)}(\tau) \tag{7.77}$$

This bivariate function is right continuous with left-hand limits and nondecreasing in each of its arguments. It also verifies

$$\lim_{\lambda \to -\infty, \tau \to -\infty} \phi_n(\lambda, \tau) = 0 \quad \text{and} \quad \lim_{\lambda \to \infty, \tau \to \infty} \phi_n(\lambda, \tau) = 1$$

Therefore, it satisfies the properties of a bivariate cumulative distribution function.

From ϕ_n we can extract precise information about the sample eigenvectors. Our goal of characterizing the behavior of sample eigenvectors would be achieved in principle by determining the asymptotic behavior of ϕ_n. This can be deduced from Theorem 7.7.2 thanks to the inversion formula for the Stieltjes transform: for all $(\lambda, \tau) \in \mathbb{R}^2$ such that ϕ_n is continuous at (λ, τ)

$$\phi_n(\lambda, \tau) = \lim_{\eta \to 0+} \frac{1}{\pi} \int_{-\infty}^{\lambda} \operatorname{Im}\left[\Phi_n^g(\xi + j\eta)\right] d\xi \tag{7.78}$$

which holds in the special case where $g = \mathbf{1}_{[-\infty, \tau)}(\lambda)$. We are now ready to state our second main result.

Theorem 7.7.3 (Ledoit and Péché (2011) [411]) Assume that conditions (H1)–(H4) hold true and let $\phi_n(\lambda, \tau)$ be defined by (7.77). Then there exists a nonrandom bivariate function ϕ such that

$$\phi_n(\lambda, \tau) \xrightarrow{\text{almost surely}} \phi(\lambda, \tau)$$

at all points of continuity of ϕ. Furthermore, when $\gamma \neq 1$, the function ϕ can be expressed as:

$$\forall (\lambda, \tau) \in \mathbb{R}^2, \quad \phi(\lambda, \tau) = \int_{-\infty}^{\lambda} \int_{-\infty}^{\tau} K(\ell, t) dH(t) dF(\ell)$$

where

$$\forall (\lambda, \tau) \in \mathbb{R}^2, \quad K(\ell, t) = \begin{cases} \frac{\gamma^{-1}\ell t}{(at-\ell)^2 + b^2 t^2} & \text{if } \ell > 0 \\ \frac{1}{(1-\gamma)\left[1 + t \lim\limits_{z \in \mathbb{C}^+ \to 0} m_G(z)\right]} & \text{if } \ell = 0 \text{ and } \gamma < 1 \\ 0 & \text{otherwise} \end{cases} \tag{7.79}$$

and a (resp. b) is the real (resp. imaginary) part of $1 - \gamma^{-1} - \gamma^{-1}\ell \tilde{m}_F(\ell)$.

(7.79) quantifies how the eigenvectors of the sample covariance matrix deviate from those of the population covariance matrix under large-dimensional asymptotics. The result is explicit as a function of $m_F(z)$.

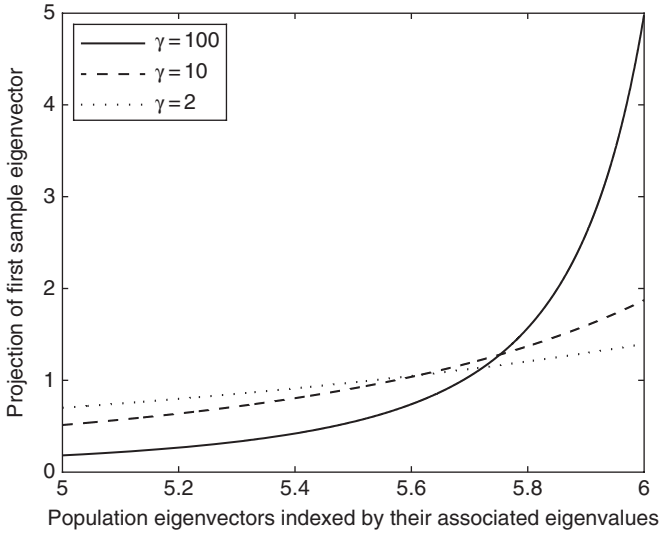


Figure 7.7 Projection of first sample eigenvector onto population eigenvectors (indexed by their their associated eigenvalues). We have taken $H' = \mathbf{1}_{[5,6)}$. Source: Reproduced from [411] with permission.

To illustrate Theorem 7.7.3, we can pick any sample eigenvector of our choosing, for example the one that corresponds to the first (i.e. largest) eigenvalue, and plot how it projects onto the true eigenvectors (indexed by their corresponding eigenvalues). The resulting graph is Figure 7.7. This is a plot of $K(\ell, t)$ as a function of t, for fixed ℓ equal to the supremum of supp (F), where supp$(\cdot)$ denotes the support of the function[1]. It is the asymptotic equivalent to plotting $n\left|\mathbf{u}_1^H \mathbf{v}_j\right|^2$ as a function of τ_j. It looks like a density because, by construction, it must integrate to one. As soon as the sample size is of the order of ten times the number of variables, we can see that the first sample eigenvector starts to deviate quite strongly from the first true eigenvectors. This should have precautionary implications for principal component analysis (PCA), where the number of variables is often so large that it is difficult to make the sample size more than ten times bigger.

7.7.3 Asymptotically Optimal Bias Correction for the Sample Eigenvalues

We now bring the two preceding results together to quantify the relationship between the sample covariance matrix and the true covariance matrix *as a whole*. This is achieved by selecting the function $g(\tau) = \tau$ in (7.76). The main problem with the sample covariance matrix is that its eigenvalues are too dispersed: the smallest ones are biased downwards, and the largest ones upwards. This is most easily visualized when the true covariance matrix is the identity, in which case the limiting spectral of sample eigenvalues F is known in closed form (see Figure 7.8).

It is reasonable to require that the estimation procedure be invariant with respect to rotation by any p-dimensional orthogonal matrix $\mathbf{W}$. If we rotate the variables by $\mathbf{W}$, then we would also ask our estimator to rotate by the same orthogonal matrix $\mathbf{W}$.

1 The support of a function is the set of points where the function is not zero, or the closure of that set.

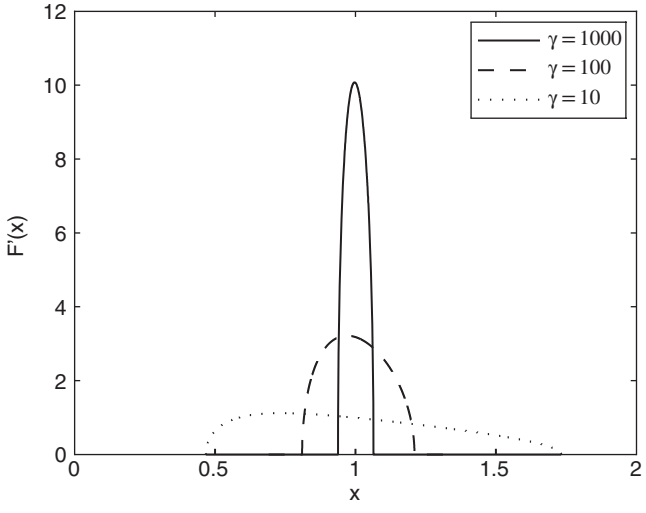


Figure 7.8 Limiting density of sample eigenvalues, in the particular case where all the eigenvalues of the true covariance matrix are equal to one. The graph shows excess dispersion of the sample eigenvalues. The formula for this plot comes from solving the Marchenko–Pastur equation for $H = \mathbf{1}_{[1,\infty)}$. Source: Reproduced from [411] with permission.

The class of orthogonally invariant estimators of the covariance matrix is constituted from all the estimators that have the same eigenvectors as the sample covariance matrix (see [412, Lemma 5.3]). Every rotation-invariant estimator of $\mathbf{\Sigma}$ is thus of the form:

$$\mathbf{U}\mathbf{D}\mathbf{U}^H, \quad \text{where } \mathbf{D} = \text{diag}\left(d_1, \ldots, d_n\right) \text{ is diagonal}$$

and where $\mathbf{U}$ is the matrix whose i-th column is the sample eigenvector $\mathbf{u}_i$.

Our objective is to find the matrix in this class that is closest to the true covariance matrix. In order to measure distance, we choose the Frobenius norm, defined as:

$$\|\mathbf{A}\|_F^2 = \text{Tr}\left(\mathbf{A}\mathbf{A}^H\right)$$

for any matrix $\mathbf{A}$. Note that the trace function $\text{Tr}(\mathbf{B})$ is linear in $\mathbf{B}$. Thus we end up with the following optimization problem:

$$\underset{\mathbf{D}}{\text{minimize}} \quad \left\|\mathbf{U}\mathbf{D}\mathbf{U}^H - \mathbf{\Sigma}\right\|_F$$

Elementary matrix algebra shows that its solution is:

$$\tilde{\mathbf{D}} = \text{diag}\left(\tilde{d}_1, \ldots, \tilde{d}_n\right) \text{ where } \forall i = 1, \ldots, n \quad \tilde{d}_i = \mathbf{u}_i^H \mathbf{\Sigma} \mathbf{u}_i$$

$\tilde{d}_i$ captures how the i-th sample eigenvector $\mathbf{u}_i$ relates to the true covariance matrix $\mathbf{\Sigma}$ as a whole.

The key object is the nondecreasing function

$$\forall x \in \mathbb{R}, \quad \Delta_n(x) = \frac{1}{n} \sum_{i=1}^{n} \tilde{d}_i \mathbf{1}_{[\lambda_i, +\infty)}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{u}_i^H \mathbf{\Sigma} \mathbf{u}_i \times \mathbf{1}_{[\lambda_i, +\infty)}(x) \tag{7.80}$$

When all the sample eigenvalues are distinct, it is straightforward to recover the $\tilde{d}_i$ from Δ_n:

$$\forall i = 1, \ldots, n \qquad \tilde{d}_i = \lim_{\varepsilon \to 0+} \frac{\Delta_n(\lambda_i + \varepsilon) - \Delta_n(\lambda_i - \varepsilon)}{F_n(\lambda_i + \varepsilon) - F_n(\lambda_i - \varepsilon)} \tag{7.81}$$

The asymptotic behavior of Δ_n can be deduced from Theorem 7.7.2 in the special case where $g(\tau) = \tau$: for all $x \in \mathbb{R}$ such that Δ_n continuous at x

$$\Delta_n(x) = \lim_{\eta \to 0+} \frac{1}{\pi} \int_{-\infty}^{x} \mathrm{Im}\left[\Phi_n^g(\xi + j\eta)\right] d\xi, \quad g(x) \equiv x \tag{7.82}$$

We are now ready to state our third main result.

Theorem 7.7.4 ([411]) Assume that conditions (H1)–(H4) hold true and let Δ_n be defined as in (7.80). There exists a nonrandom function Δ defined over $\mathbb{R}$ such that $\Delta_n(x)$ converges almost surely to $\Delta(x)$ for all $x \in \mathbb{R} - \{0\}$. If in addition $\gamma \neq 1$, then Δ can be expressed as:

$$\forall x \in \mathbb{R}, \quad \Delta(x) = \int_{-\infty}^{x} \psi\{\lambda\} \, dF(\lambda)$$

where

$$\forall x \in \mathbb{R}, \quad \psi(\lambda) = \begin{cases} \frac{\lambda}{\left|1 - \gamma^{-1} - \gamma^{-1}\lambda \tilde{m}_F(\lambda)\right|^2} & \text{if } \lambda > 0 \\ \frac{\gamma}{(1-\gamma) \lim_{z \in \mathbb{C}^+ \to 0} \tilde{m}_G(z)} & \text{if } \lambda = 0 \text{ and } \gamma < 1 \\ 0 & \text{otherwise} \end{cases} \tag{7.83}$$

By (7.81) the asymptotic quantity that corresponds to $\tilde{d}_i = \mathbf{u}_i^H \boldsymbol{\Sigma} \mathbf{u}_i$ is $\psi\{\lambda\}$, provided that λ corresponds to λ_i. Therefore, the way to get closest to the true covariance matrix (according to the Frobenius norm) would be to divide each sample eigenvalue λ_i by the correction factor

$$\left|1 - \gamma^{-1} - \gamma^{-1}\lambda \tilde{m}_F(\lambda)\right|^2$$

This is what we call the optimal nonlinear shrinkage formula or asymptotically optimal bias correction. Figure 7.9 shows how much it differs from Ledoit–Wolf [402] optimal linear shrinkage formula. In addition, when $\gamma < 1$, the sample eigenvalues equal to zero need to be replaced by

$$\psi(0) = \frac{\gamma}{(1-\gamma)\tilde{m}_G(0)}$$

For each set of simulations, we computed the percentage relative improvement in average loss (PRIAL). The PRIAL of an estimator $\mathbf{M}$ of $\boldsymbol{\Sigma}$ is defined as

$$PRIAM(\mathbf{M}) = 100 \times \left[1 - \frac{\left\|\mathbf{M} - \mathbf{U}\tilde{\mathbf{D}}\mathbf{U}^H\right\|_F^2}{\left\|\mathbf{S} - \mathbf{U}\tilde{\mathbf{D}}\mathbf{U}^H\right\|_F^2}\right]$$

By construction, the PRIAL of the sample covariance matrix $\mathbf{S}$ (resp. of $\mathbf{U}\tilde{\mathbf{D}}\mathbf{U}^H$) is 0% (resp. 100%), meaning no improvement (resp. meaning maximum attainable

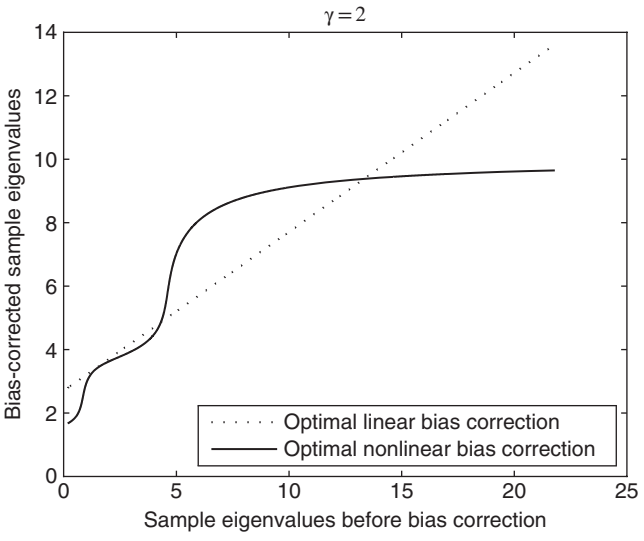


Figure 7.9 Comparison of the optimal linear versus nonlinear bias correction formula. The distribution of true eigenvalues *H* places 20% mass at 1, 40% mass at 3 and 40% mass at 10. Source: Reproduced from [411] with permission.

improvement). As shown in Figure 7.10, even with a modest sample size like $N = 40$, we already get 95% of the maximum possible improvement.

A similar formula can be obtained for the purpose of estimating the *inverse* of the true covariance matrix, $\mathbf{\Sigma}^{-1}$. To this aim, we set $g(\tau) = 1/\tau$ in (7.76).

7.7.4 Estimating Precision Matrices

Suppose observations $\mathbf{x}_1, \ldots, \mathbf{x}_N$ independently from a multivariate model

$$\mathbf{x}_i = \mathbf{\Sigma}_n^{1/2}\mathbf{y}_i + \mu_0, \quad i = 1, \ldots, N \tag{7.84}$$

where μ_0 is a n-dimensional constant vector and $\mathbf{\Sigma}_n$ is a $n \times n$ positive definite matrix, which acts as the true covariance matrix. Here

$$\mathbf{Y} = \left(\mathbf{y}_1, \ldots, \mathbf{y}_N\right) = \left(Y_{ij}\right)_{n\times N}$$

and $Y_{ij}, i, j = 1, 2, \ldots$ are real independent and identically distributed (i.i.d.) random variables with common mean zero and unit variance. In multivariate analysis, estimation of the covariance matrix $\mathbf{\Sigma}_n$ and precision matrix $\mathbf{\Omega}_n = \mathbf{\Sigma}_n^{-1}$ is an important problem. Given the samples $\mathbf{x}_1, \ldots, \mathbf{x}_N$, the usual estimation of $\mathbf{\Sigma}_n$ is the sample covariance matrix which is defined as

$$\mathbf{S}_N = \frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{x}_i - \overline{\mathbf{x}}\right)\left(\mathbf{x}_i - \overline{\mathbf{x}}\right)^T \tag{7.85}$$

where $\overline{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$ and the superscript T denotes the transpose of a matrix or vector. Naturally, in many areas of statistical analysis, $\mathbf{S}_N^{-1}$ acts as the common estimator of $\mathbf{\Omega}_n = \mathbf{\Sigma}_n^{-1}$.

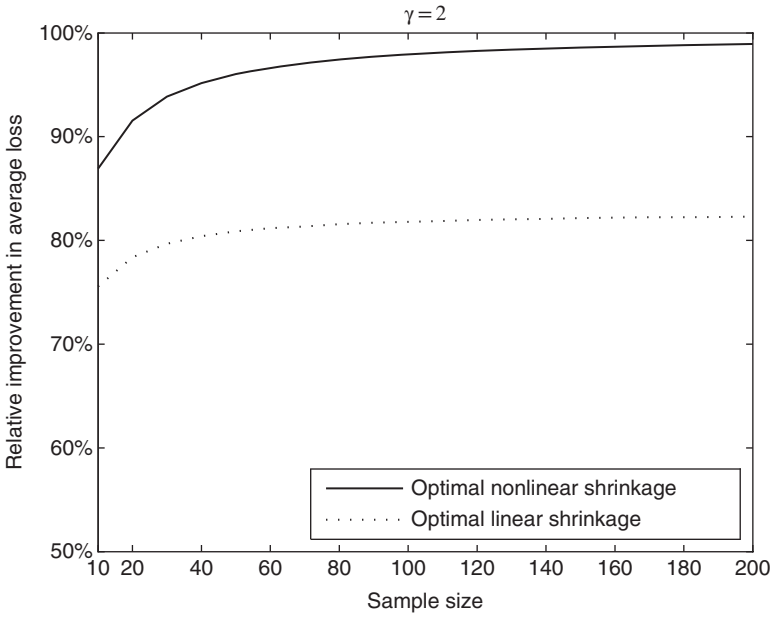


Figure 7.10 Percentage relative improvement in average loss (PRIAL) from applying the optimal nonlinear shrinkage formula to the sample eigenvalues. The solid line shows the PRIAL obtained by dividing the i-th sample eigenvalue by the correction factor $\left|1-\gamma^{-1}-\gamma^{-1}\lambda_i\tilde{m}_F\left(\lambda_i\right)\right|^2$, as a function of sample size. The dotted line shows the PRIAL of the linear shrinkage estimator first proposed in [402]. For each sample size we ran 10 000 Monte Carlo simulations. As in Figure 7.9, we used $\gamma = 2$ and the distribution of true eigenvalues H placing 20% mass at 1, 40% mass at 3 and 40% mass at 10. Source: Reproduced from [411] with permission.

In classic statistics where the dimension n is fixed and the sample size $N \to \infty$, $\mathbf{S}_N^{-1}$ is a good estimator for $\mathbf{\Omega}_n = \mathbf{\Sigma}^{-1}$. In the large dimensional data setting where the data dimension n is large compared to the sample size N, the usual estimator, which simply takes the inverse of the sample covariance matrix, has two disadvantages. First, $\mathbf{S}_N^{-1}$ is singular if $n > N$, which means we cannot obtain a stable estimator for $\mathbf{\Omega}_n$. Secondly, even if $n < N$, $\mathbf{S}_N^{-1}$ as the estimator of $\mathbf{\Omega}_n$ is known to perform poorly. For example, if $n/N \to \gamma \in (0, 1)$, by Remark 2 of Pan and Zhou (2011) [413], we have

$$\mathrm{Tr}\left(\mathbf{\Sigma}\mathbf{S}_N^{-1} - \mathbf{I}_n\right)^2 \xrightarrow{p} \frac{\gamma\left(1+\gamma-\gamma^2\right)}{(1-\gamma)^3} \tag{7.86}$$

which shows that the estimation error is dramatically large, especially when $\gamma \to 1$. Here $\mathbf{I}_n$ is the $n \times n$ identity matrix.

Mestre and Lagunas (2006) [414] suggested

$$(1 - n/N)\,\mathbf{S}_N^{-1}$$

to estimate $\mathbf{\Omega}_n = \mathbf{\Sigma}^{-1}$. Although the shrinking estimators proposed in [402] and successive works are invertible and more accurate than sample covariance matrix $\mathbf{S}_N$ to estimate $\mathbf{\Sigma}_n$, their inverses are usually not the best for $\mathbf{\Sigma}_n$ among the combinations $\mathbf{S}_N$ and $\mathbf{I}_n$. Moreover, the methods in [414] and [415] are only applicable for $n < N$.

Following [416], motivated by Ledoit and Peche (2011), we will study the asymptotic properties of the matrix

$$\mathbf{\Sigma}_n^{1/2}\left(\mathbf{S}_N + \lambda\mathbf{I}_n\right)^{-1}\mathbf{\Sigma}_n^{1/2}$$

and its relation with $\left(\mathbf{S}_N + \lambda\mathbf{I}_n\right)^{-1}$. Based on these limiting results, [416] propose an optimal linear combination of $\mathbf{S}_N$ and $\mathbf{I}_n$ under the quadratic loss function

$$\frac{1}{n}\operatorname{Tr}\left(\mathbf{\Sigma}_n\left(\lambda_1\mathbf{S}_N + \lambda_2\mathbf{I}_n\right)^{-1} - \mathbf{I}_n\right)^2$$

The new estimation is nonparametric without assuming a specific parameter distribution for the data and also there is no prior information about the structure of the population covariance matrix. The new estimator has no restriction on $n < N$ and is applicable for $n \geq N$. Even if $n < N$, the new estimator always dominates the standard $\mathbf{S}_N^{-1}$ and $(1 - n/N)\,\mathbf{S}_N^{-1}$ proposed in [414]. It also performs comparably with the nonlinear shrinkage estimator in [415].

Here, we make the following assumptions:

A1) $n, N \to \infty$ such that $n/N = \gamma \in (0, \infty)$ and and the fourth moment of Y_{ij} is bounded;
A2) The extreme eigenvalues of $\mathbf{\Sigma}_n$ is uniformly bounded so that there are constants c_1, c_2 satisfying $c_1 \leqslant \lambda_{\min}\left(\mathbf{\Sigma}_n\right) \leqslant \lambda_{\max}\left(\mathbf{\Sigma}_n\right) \leqslant c_2$ and $F_{\mathbf{\Sigma}_n}$ tends to a nonrandom probability distribution H.

Now, we can introduce a lemma about the limiting spectral distribution of $\mathbf{\Sigma}_n^{1/2}\left(\mathbf{S}_N + \lambda\mathbf{I}_n\right)^{-1}\mathbf{\Sigma}_n^{1/2}$.

Theorem 7.7.5 (Wang, Pan and Cao (2012) [416]) Under the conditions of A1 and A2, as $N \to \infty$, $F_{\mathbf{\Sigma}_n^{1/2}\left(\mathbf{S}_N+\lambda\mathbf{I}_n\right)^{-1}\mathbf{\Sigma}_n^{1/2}}$ converges almost certainly to a nonrandom distribution F, whose Stieltjes transform $m(z)$ satisfies

$$m(z) = \int \frac{1}{\frac{\lambda}{t} - z + \frac{1}{1+\gamma m(z)}} dH(t) \tag{7.87}$$

where $\lambda > 0$ and $z \in \mathbb{C}^+ = \{z \in \mathbb{C}, \operatorname{Im}(z) > 0\}$.

The result of Theorem 7.7.5 also can be derived from Theorem 1.2 in [411] where the 12th moment is needed.

From [175], we know the Stieltjes transform $m_0(z)$ of the limiting spectral distribution of $\mathbf{S}_N$ is the solution to the following equation

$$m_0(z) = \int \frac{1}{t\left(1 - \gamma - \gamma z m_0(z)\right) - z} dH(t) \tag{7.88}$$

For more analytic behaviors of (7.88), see (1995) [222].

Now, we can study the relationships between $\mathbf{\Sigma}_n^{1/2}\left(\mathbf{S}_N + \lambda\mathbf{I}_n\right)^{-1}\mathbf{\Sigma}_n^{1/2}$ and $\left(\mathbf{S}_N + \lambda\mathbf{I}_n\right)^{-1}$.

Theorem 7.7.6 ([416]) When $\lambda > 0$, under the conditions of A1 and A2, as $N \to \infty$, we almost certainly have

$$\frac{1}{n}\operatorname{Tr}\left[\mathbf{\Sigma}_n^{1/2}\left(\mathbf{S}_N + \lambda\mathbf{I}_n\right)^{-1}\mathbf{\Sigma}_n^{1/2}\right] \to R_1(\lambda)$$
$$\frac{1}{n}\operatorname{Tr}\left[\mathbf{\Sigma}_n^{1/2}\left(\mathbf{S}_N + \lambda\mathbf{I}_n\right)^{-1}\mathbf{\Sigma}_n^{1/2}\right]^2 \to R_2(\lambda)$$

and

$$\frac{1}{n}\operatorname{Tr}\left[\left(\mathbf{S}_N+\lambda\mathbf{I}_n\right)^{-1}\right]\to m_0\left(-\lambda\right)$$

Moreover

$$R_1\left(\lambda\right)=\frac{1-\lambda m_0\left(-\lambda\right)}{1-\gamma\left(1-\lambda m_0\left(-\lambda\right)\right)}$$

$$R_2\left(\lambda\right)=\frac{1-\lambda m_0\left(-\lambda\right)}{\left[1-\gamma\left(1-\lambda m_0\left(-\lambda\right)\right)\right]^3}-\frac{\lambda m_0\left(-\lambda\right)-\lambda^2 m_0^{'}\left(-\lambda\right)}{\left[1-\gamma\left(1-\lambda m_0\left(-\lambda\right)\right)\right]^4}$$

Here, almost surely, $\frac{1}{n}\operatorname{Tr}\left[\left(\mathbf{S}_N+\lambda\mathbf{I}_n\right)^{-2}\right]\to m'\left(-\lambda\right)=\frac{dm(z)}{dz}\Big|_{z=-\lambda}$

In applications, we cannot derive the statistics $\frac{1}{n}\operatorname{Tr}\left[\boldsymbol{\Sigma}_n^{1/2}\left(\mathbf{S}_N+\lambda\mathbf{I}_n\right)^{-1}\boldsymbol{\Sigma}_n^{1/2}\right]^k, k=1,2$ directly because only $\mathbf{S}_N$ is known. Theorem 7.7.6 provided a theoretical method to estimate the statistics from the sample covariance matrix. In [417], the authors derived a similar result to that of Theorem 2 under Gaussian assumptions and here no distribution assumptions were needed.

About $m_0(-\lambda)$ and (7.88), we have the following result. In Theorem 7.7.6, $m_0(-\lambda)$ is the unique solution of the equation

$$m(-\lambda)=\int\frac{1}{t\left(1-\gamma-\gamma\lambda m\left(-\lambda\right)\right)+\lambda}dH(t) \tag{7.89}$$

where $1-\gamma-\gamma\lambda m\left(-\lambda\right)\geqslant 0$. Assuming $\boldsymbol{\Sigma}_n=\mathbf{I}_n$, two solutions of (7.89) can be written out as follows

$$m^{(1)}(-\lambda)=\frac{1}{2\gamma\lambda}\left(-\left(1-\gamma+\lambda\right)+\sqrt{\left(1-\gamma+\lambda\right)^2+4\gamma\lambda}\right)$$

$$m^{(2)}(-\lambda)=\frac{1}{2\gamma\lambda}\left(-\left(1-\gamma+\lambda\right)-\sqrt{\left(1-\gamma+\lambda\right)^2+4\gamma\lambda}\right)$$

Optimal Estimator

To estimate $\boldsymbol{\Omega}_n=\boldsymbol{\Sigma}_n^{-1}$, we consider a class of estimators such as $\hat{\boldsymbol{\Omega}}_n=\alpha\left(\mathbf{S}_N+\beta\mathbf{I}_n\right)^{-1}$. By Theorem 7.7.6, with probability 1,

$$\begin{aligned}\frac{1}{n}\operatorname{Tr}\left(\boldsymbol{\Sigma}\hat{\boldsymbol{\Omega}}_n-\mathbf{I}_n\right)^2&\to\alpha^2R_2\left(\beta\right)-2\alpha R_1\left(\beta\right)+1\\&=R_2\left(\beta\right)\left(\alpha-\frac{R_1(\beta)}{R_2(\beta)}\right)^2+1-\frac{\left(R_1(\beta)\right)^2}{R_2(\beta)}.\end{aligned} \tag{7.90}$$

Therefore, to minimize the loss function (7.90), α should satisfy $\alpha=\frac{R_1(\beta)}{R_2(\beta)}$ and the corresponding loss is

$$L\left(\beta\right)=1-\frac{\left(R_1\left(\beta\right)\right)^2}{R_2\left(\beta\right)} \tag{7.91}$$

Intuitively, β should minimize $L\left(\beta\right)$. About the optimal loss $L_0=\min_{\beta>0}L\left(\beta\right)$, we have the following results.

Theorem 7.7.7 ([416]) When $\gamma < 1$, writing

$$L_H(y) = 1 - \left(\int \frac{t}{t+y} dH(t)\right)^2 \left(\frac{1}{\int \frac{t^2}{(t+y)^2} dH(t)} - \gamma\right), y \geqslant 0$$

we have $L_0 = \min_{y>0} L_H(y)$. Moreover,
I.$\mathbf{\Sigma}_n = \sigma^2 \mathbf{I}_n$, which means $H(x)$ is a degenerate distribution at σ^2, the optimal loss is $L_0 = 0$ and $\mathbf{\Omega}_n^\star = \sigma^{-2}\mathbf{I}_n$.
II. For general distribution $H(x)$, $L_H(y)$ achieves its global minimum values L_0 at $y^\star$ satisfying

$$\frac{f_1(y^\star) f_3(y^\star) - f_2(y^\star) f_2(y^\star)}{f_2(y^\star) f_2(y^\star) \left(f_1(y^\star) - f_2(y^\star)\right)} = \gamma \tag{7.92}$$

where $f_k(x) = \int \left(\frac{t}{t+y}\right)^k dH(t)$. Moreover, $\beta^\star$ satisfies the equation $y^\star = \frac{\beta^\star}{1-\gamma\left(1-\beta^\star m_0(-\beta^\star)\right)}$ and $\alpha^\star = \frac{R_1(\beta^\star)}{R_2(\beta^\star)}$

For $\gamma > 1$, we can also obtain a similar result from the proofs of Theorem 7.7.7 and therefore will not pursue it due to limited space. The following corollary gives a data-driven method to estimate the best β.

Corollary 7.7.8 (Wang, Pan and Cao (2012) [416]) Under the assumptions of Theorem 7.7.6 and writing $\hat{\gamma} = n/N$

$$a_1(\lambda) = 1 - \frac{1}{n} \operatorname{Tr}\left(\lambda_1 \mathbf{S}_N + \lambda_2 \mathbf{I}_n\right)^{-1}$$
$$a_2(\lambda) = \frac{1}{n} \operatorname{Tr}\left(\frac{1}{\lambda}\mathbf{S}_N + \mathbf{I}_n\right)^{-1} - \frac{1}{n} \operatorname{Tr}\left(\frac{1}{\lambda}\mathbf{S}_N + \mathbf{I}_n\right)^{-2}$$

and

$$\hat{R}_1(\lambda) = \frac{a_1(\lambda)}{1 - \hat{\gamma} a_1(\lambda)}$$
$$\hat{R}_2(\lambda) = \frac{a_1(\lambda)}{\left(1 - \hat{\gamma} a_1(\lambda)\right)^3} - \frac{a_2(\lambda)}{\left(1 - \hat{\gamma} a_1(\lambda)\right)^4}$$

almost surely, as $N \to \infty$

$$\hat{R}_1(\lambda) \to R_1(\lambda)$$
$$\hat{R}_2(\lambda) \to R_2(\lambda)$$

By Corollary 7.7.8 and continuous mapping theorem, with probability 1

$$\hat{L}(\beta) := 1 - \frac{\left(\hat{R}_1(\beta)\right)^2}{\hat{R}_2(\beta)} \to L(\beta)$$

Therefore, for a real data or a sample covariance matrix $\mathbf{S}_N$, we can use a numerical algorithm such as the Newton–Raphson method to find the optimal $\beta^\star$ from $\hat{L}(\beta)$.

Compared with existing methods, the new estimator

$$\mathbf{\Omega}_n^\star = \alpha^\star \left(\mathbf{S}_N + \beta^\star \mathbf{I}_n\right)^{-1}$$

has the following properties. First, when $\gamma < 1$, by Remark 2 of Pan and Zhou (2011) [413]

$$\frac{1}{n} \operatorname{Tr}\left((1-\gamma)\,\mathbf{\Sigma}_n \mathbf{S}_N^{-1} - \mathbf{I}_n\right)^2 \xrightarrow{p} \frac{\gamma}{1-\gamma} \tag{7.93}$$

which is the loss of estimator in [414]. By proofs of Theorem 7.7.7, we know the optimal loss of our estimator is $L_0 < L_H(0) = \gamma$. Together with formula (7.86), it is shown that when $\gamma < 1$, the new estimator will always dominate the standard $\mathbf{S}_N^{-1}$ and $(1 - n/N)\,\mathbf{S}_N^{-1}$ proposed in [414]. Secondly, when $\gamma \geq 1$, we can still use the new estimator while the estimators based on $\mathbf{S}_N^{-1}$ or the nonlinear shrinkage estimator based on the eigenvalues of $\mathbf{S}_N$ in [415] are not applicable any more.

7.8 A General Class of Random Matrices

Motivated by practical applications in Section 8.2, we address a general class of random matrices that was first studied by Rubio and Mestre [418].

Let $\mathbf{X}$ be an $M \times N$ complex random matrix with i.i.d. entries having mean zero and variance $1/N$ with finite $8 + \varepsilon$ moment ($\varepsilon > 0$). Furthermore, consider an $M \times M$ Hermitian non-negative definite matrix $\mathbf{R}$ and and its non-negative definite square-root $\mathbf{R}^{1/2}$. Then, the matrix $\mathbf{S} = \mathbf{R}^{1/2}\mathbf{X}\mathbf{T}\mathbf{X}^H\mathbf{R}^{1/2}$ be viewed as a sample covariance matrix constructed using the N columns of the data matrix $\mathbf{R}^{1/2}\mathbf{X}$, namely having true population covariance matrix $\mathbf{R}$. Besides, consider an $N \times N$ diagonal matrix $\mathbf{T}$ with real non-negative entries. The matrix $\mathbf{R}^{1/2}\mathbf{X}\mathbf{T}\mathbf{X}^H\mathbf{R}^{1/2}$ can be interpreted as a sample covariance matrix obtained by weighting the previous multivariate samples with the entries of $\mathbf{T}$.

We are interested in the asymptotic behavior of certain spectral functions of the random matrix model

$$\mathbf{B} = \mathbf{A} + \mathbf{R}^{1/2}\mathbf{X}\mathbf{T}\mathbf{X}^H\mathbf{R}^{1/2} \tag{7.94}$$

where $\mathbf{A}, R$ and $\mathbf{T}$ are Hermitian non-negative definite matrices, such that $\mathbf{R}$ and $\mathbf{T}$ have bounded spectral norm with being diagonal, and $\mathbf{R}^{1/2}$ is the non-negative definite square-root of $\mathbf{R}$. Under some assumptions on the moments of the entries of $\mathbf{X}$, it is proved that, for any matrix $\mathbf{\Phi}$ with bounded trace norm and for each complex z outside the positive real line

$$\left|\operatorname{Tr}\left[\mathbf{\Phi}\left(\mathbf{B} - z\mathbf{I}_M\right)^{-1}\right] - g_M(z)\right| \to 0 \quad \text{almost certainly} \tag{7.95}$$

as $M, N \to \infty$ at the same rate, where $g_M(z)$ is deterministic and solely depends on $\mathbf{\Phi}, \mathbf{A}, \mathbf{R}$ and $\mathbf{T}$. (7.95) can be particularized to the study of the limiting behavior of the Stieltjes transform as well as the eigenvectors of the random matrix model $\mathbf{B}$.

Example 7.8.1 (estimation of large dimensional precision matrix) Let $\mathbf{\Sigma}_N$ stand for the true covariance matrix and $\mathbf{S}_N$ denote the corresponding sample covariance matrix.

The pairs $(\tau_i, \mathbf{v}_i)$ for $i = 1, \ldots, n$ denote the collection of eigenvalues and the corresponding orthonormal eigenvectors of the covariance matrix $\mathbf{\Sigma}_N$. $H_N(t)$ is the empirical distribution function of the eigenvalues of $\mathbf{\Sigma}_n$:

$$H_N = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{\tau_i < t\}}(t), \quad \forall t \in \mathbb{R}$$

where $\mathbf{1}_{\{\cdot\}}$ is the indicator function of the set. Let $\mathbf{X}_N$ be a $n \times N$ matrix which consists of independent and identically distributed (i.i.d.) real random variables with zero mean and unit variance. The observation matrix is defined as

$$\mathbf{Y}_N = \mathbf{\Sigma}_N^{1/2}\mathbf{X}_N$$

The pairs $\left(\lambda_i, \mathbf{u}_i\right)$ for $i = 1, \ldots, n$ are the eigenvalues and the corresponding orthonormal eigenvectors of the sample covariance matrix

$$\mathbf{S}_N = \frac{1}{N}\mathbf{Y}_N\mathbf{Y}_N^H = \frac{1}{N}\mathbf{\Sigma}_N^{1/2}\mathbf{X}_N\mathbf{X}_N^H\mathbf{\Sigma}_N^{1/2}$$

Similarly, the empirical distribution function of the eigenvalues of the sample covariance matrix $\mathbf{S}_N$ is defined as

$$F_N(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{\lambda_i < \lambda\}}(\lambda), \quad \forall \lambda \in \mathbb{R}$$

The main assumptions are made here:

(A1) The true population covariance matrix $\mathbf{\Sigma}_N$ is a nonrandom n-dimensional positive definite matrix.
(A2) Only the matrix $\mathbf{Y}_N$ is observable. We know neither $\mathbf{X}_N$ nor $\mathbf{\Sigma}_N$ itself.
(A3) We assume that $H_N(t)$ converges to a limit $H(t)$ at all points of continuity of $H(t)$.
(A4) The elements of the matrix $\mathbf{X}_N$ have uniformly bounded $4 + \varepsilon, \varepsilon > 0$ moments.
(A5) For all N large enough there exists the compact interval $[h_0, h_1]$ in $(0, +\infty)$ which contains the support of H_N.

All of these assumptions are quite general and are satisfied in many practical situations. The assumptions (A1)–(A3) are essential to prove the Marchenko–Pastur equation.

Assume that (A1), (A2), (A4), (A5) hold and additionally some nonrandom matrix $\mathbf{\Phi}$ has uniformly bounded trace norm at infinity then for $n/N \to c > 0$ as $N \to \infty$

$$\left|\mathrm{Tr}\left[\mathbf{\Phi}\left(\mathbf{S}_N - z\mathbf{I}_N\right)^{-1}\right] - \mathrm{Tr}\left[\mathbf{\Phi}\left(x(z)\,\mathbf{\Sigma}_N - z\mathbf{I}_N\right)^{-1}\right]\right| \to 0 \quad \text{almost certainly,}$$

where $x(z)$ is a unique solution in $\mathbb{C}^+$ of the following equation

$$\frac{1 - x(z)}{x(z)} = \frac{c}{p}\,\mathrm{Tr}\left(x(z)\mathbf{I}_N - z\mathbf{\Sigma}_N^{-1}\right)^{-1}$$

See [419,420] for proofs. □

Example 7.8.2 (array signal processing) Consider a sensor network with n sensors observing N successive snapshots of K source signals. Most statistical inference methods for array processing assume an array of size N fixed and a number of snapshots

N large. In addition, many works are based on the assumption of a white-noise model. These two assumptions are increasingly less realistic in modern systems where n and N are usually both large, and where the noise data can be correlated either across successive observations or across the sensor antennas. It is natural to assume the asymptotic regime denoted by

$$N \to \infty, n \to \infty \text{ but } n/N \to c > 0$$

The number of transmitting sources K is fixed as $N \to \infty$.

In this section, apart from n and N, all parameters including K are unknown. In particular, the noise spatial or temporal correlations are unknown. The angle taken in this section to perform statistical inference on the signals is based on the spectral analysis of the empirical covariance matrix of the received signals.

Consider K source signals received by an array of n sensors during N time slots. The received signal $\mathbf{y}_t \in \mathbb{C}^{n\times 1}$ at time t is given by

$$\mathbf{y}_t = \sum_{i=1}^{K} \sqrt{P_k}\mathbf{a}_N\left(\theta_k\right)\mathbf{s}_{k,t} + \mathbf{v}_t$$

where P_k is the power of source k, $\theta_k \in \left[-\pi/2, \pi/2\right]$ is its angle of arrival (different for each k), $\mathbf{a}_N \in \mathbb{C}^{n\times 1}$ is the steering vector defined in the classical uniform linear array model as

$$\mathbf{a}_N\left(\theta_k\right) = \frac{1}{\sqrt{n}}\left[1, e^{-j2\pi d \sin\theta_k}, \ldots, e^{-j2\pi d(n-1)\sin\theta_k}\right]^T$$

with d a positive constant. We can rewrite the input-output relationship by concatenating T successive signal realizations into the matrix

$$\mathbf{Y}_N = \mathbf{H}_N\mathbf{P}^{1/2}\mathbf{S}_N^H + \mathbf{V}_N$$

where

$$\mathbf{Y}_N = \left[\mathbf{y}_1, \ldots, \mathbf{Y}_N\right], \mathbf{H}_N = \left[\mathbf{a}_N\left(\theta_1\right), \ldots, \mathbf{a}_N\left(\theta_K\right)\right],$$
$$\mathbf{P} = diag\left(P_1, \ldots, P_k\right), \mathbf{S}_N = \frac{1}{\sqrt{N}}\left[s_{t,k}^*\right]_{t,k=1}^{N,K}$$

with $s_{t,k}$ random i.i.d. with zero mean, unit variance, and finite eighth order moment, and $\mathbf{V}_N = \left[\mathbf{v}_1, \ldots, \mathbf{v}_N\right]$. We assume that the noise is temporally correlated, i.e., the columns of $\mathbf{V}_N$ are not independent. Although this is not a necessary condition for the validity of the results here, we assume that the noise model is a causal stationary autoregressive moving average (ARMA). process. For more details, we refer to [421, 422]. □

We illustrate a distributed system, in Figure 7.11, with a large number of sensors that are connected with cloud storage and computing through fiber cables or wire-line cables. By sensors we mean in a generic sense; examples include wireless communications for smart meters, PMUs, and a large array of antennas. We use this model to highlight the aspects of large datasets, illustrating how large-dimensional random matrices arise naturally in these situations. The receiver end of this is composed of a large number of sensors and is *unaware* of the noise pattern.

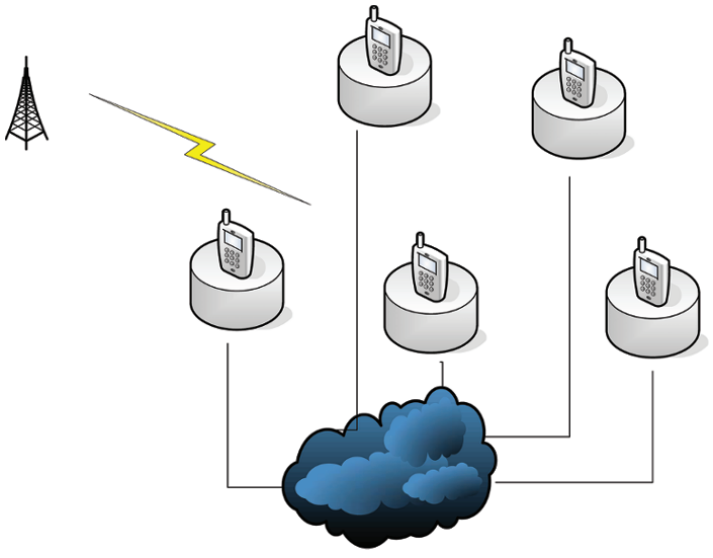


Figure 7.11 A distributed system with a large number of sensors.

Consider a very general information-plus-noise transmission model with multivariate output $\mathbf{y}_t \in \mathbb{C}^N$ at time t

$$\mathbf{y}_t = \mathbf{H}\mathbf{x}_t + \mathbf{v}_t \tag{7.96}$$

where $\mathbf{x}_t \in \mathbb{C}^K$ is the vector of transmitted symbols at time t, $\mathbf{H} \in \mathbb{C}^{N\times K}$ is the *linear* communication medium, and $\mathbf{v}_t \in \mathbb{C}^N$ the noise experienced by the receiver at time t. We assume the observation of T (not necessarily independent) vector samples $\mathbf{y}_1, \ldots, \mathbf{y}_T$ of the vector process $\mathbf{y}_t$. We use the following notation

$$\mathbf{Y}_T = \frac{1}{\sqrt{T}}\left[\mathbf{y}_1, \ldots, \mathbf{y}_T\right], \mathbf{H} = \left[\mathbf{h}_1, \ldots, \mathbf{h}_K\right],$$
$$\mathbf{X}_T = \frac{1}{\sqrt{T}}\left[\mathbf{x}_1, \ldots, \mathbf{x}_T\right], \mathbf{V}_T = \frac{1}{\sqrt{T}}\left[\mathbf{v}_1, \ldots, \mathbf{v}_T\right]$$

It is often of interest to retrieve information from the individual $\mathbf{h}_k$ vectors. In wireless communications, these represent channel beams that the receiver may want to identify in order to decode the entries. In array processing, they stand for steering vectors parameterized by the angle-of-arrival of the source signals.

The standard inference approaches in the literature often rely on two strong assumptions: (i) T is large compared to N and (ii) the statistics of $\mathbf{v}_t$ are partially or perfectly known due to independent (information-free) observations of the process $\mathbf{v}_t$. In this section we revisit these methods by surveying alternative algorithms in the recent literature to perform eigen inference for the model (7.96) accounting for the aforementioned limitations (i) and (ii).

We assume a set of reasonable conditions:

- $N \to \infty, N/T \to c > 0$, and K is constant. This allows for $\mathbf{Y}_T\mathbf{Y}_T^H$ to be seen as a small rank perturbation of $\mathbf{V}_T\mathbf{V}_T^H$.

- $\mathbf{V}_T = \mathbf{W}_T \mathbf{\Sigma}_T^{1/2}$ (i.e. white in space, correlated in time), where $\mathbf{W}_T \in \mathbb{C}^{N \times T}$ is standard complex Gaussian and $\mathbf{\Sigma}_T$ is a deterministic *unknown* Hermitian non-negative, or $\mathbf{V}_T = \mathbf{\Sigma}_T^{1/2} \mathbf{W}_T$ (i.e. white in time, correlated in space).[2]
- As $N/T \to c$, the eigenvalues of $\mathbf{V}_T \mathbf{V}_T^H$ tend to cluster in a compact interval. This assumption is satisfied by most noise models used in practice, for example auto-regressive moving average (ARMA) noise processes.
- The source signals in $\mathbf{x}_t$ are random, independent and identically distributed (i.i.d.), even though this assumption can be relaxed in many cases.

Under these assumptions, [422] shows that a maximum of K isolated eigenvalues of signal-plus-noise sample covariance matrix $\mathbf{Y}_T \mathbf{Y}_T^H$ can be found for all large N, T beyond the right edge of the limiting eigenvalue distribution support of noise only sample covariance matrix $\mathbf{V}_T \mathbf{V}_T^H$. This phenomenon is at the origin of the detection and estimation procedures. It is shown in [422] that the isolated eigenvalues of T can be *uniquely* mapped to $\mathbf{Y}_T \mathbf{Y}_T^H$ individual signal sources. The presence of these eigenvalues will be used to detect signal sources as well as to estimate their number K while their values will be exploited to estimate the source powers. The associated eigenvectors will then be used to retrieve information on the vectors $\mathbf{h}_k$.

An exemplary application of these methods to array processing is then studied in greater detail, leading in particular to a novel MUSIC-like algorithm [423] assuming *unknown noise covariance*.

7.8.1 Massive MIMO System

Section 7.8.1 is mainly based on [424], with some material drawing from other work.

The proper evaluation of the achievable rate in the MIMO setting relies on the knowledge of the transmit-receive channel as well as of the interference pattern. It is fundamental for a receiver to be able to infer these achievable rates in a short sensing period, hence it should be extremely fast. We present a novel estimator for fast estimation of the MIMO mutual information in the presence of *unknown interference* in the case where the number of available observations is of the same order as the number of receive antennas. Novel algorithms, based on large-dimensional random matrix theory, will not perform the (usually time-consuming) evaluation of the covariance matrix of the interference.

Consider a wireless communication channel $\mathbf{H}_t \in \mathbb{C}^{N \times n_0}$ between a transmitter equipped with n_0 antennas and a receiver equipped with N antennas, the latter being exposed to interfering signals. The objective of the receiver is to evaluate the mutual information from this link during a *sensing period* assuming $\mathbf{H}_t$ known at all time. For this, we assume a block-fading scenario and denote by $T \geq 1$ the number of channel coherence intervals (or time slots) allocated for sensing. In other words, we suppose that, within each channel coherence interval $t \in \{1, \ldots, T\}$, $\mathbf{H}_t$ is *deterministic* and *constant*. We also denote by M the number of channel uses employed for sensing during each time slot (M times the channel use duration is therefore less than the channel

2 Assuming the general correlated noise in both time and space would lead to too much indetermination and is so far too difficult to address.

coherence time). The M concatenated signal vectors received in slot t are gathered in the matrix $\overline{\mathbf{Y}}_t \in \mathbb{C}^{N\times M}$ defined as

$$\overline{\mathbf{Y}}_t = \mathbf{H}_t\mathbf{X}_{t,0} + \overline{\mathbf{W}}_t$$

where $\mathbf{X}_{t,0}$ is the concatenated matrix of the transmitted signals and $\overline{\mathbf{W}}_t \in \mathbb{C}^{N\times M}$ stands for the concatenated interference vectors.

Since $\overline{\mathbf{W}}_t$ is not necessarily a white noise matrix in the present scenario, we write

$$\overline{\mathbf{W}}_t = \mathbf{G}_t\mathbf{W}_t$$

where $\mathbf{G}_t \in \mathbb{C}^{N\times n}$ is such that $\mathbf{G}_t\mathbf{G}_t^H \in \mathbb{C}^{N\times N}$ is the deterministic matrix of the noise variance during slot t, while $\mathbf{W}_t \in \mathbb{C}^{n\times M}$ is a matrix filled with independent entries with zero mean and unit variance. That is, we assume that the interference is stationary during the coherence time of $\mathbf{H}_t$, which is a reasonable assumption in practical scenarios, as illustrated in Figure 7.12. We assume that perfect decoding of $\mathbf{X}_{t,0}$ (possibly transmitted at low rate or not transmitted at all) is achieved during the sensing period. If so, since $\mathbf{H}_t$ is assumed perfectly known, the residual signal to which the receiver has access is given by the standard MIMO model

$$\mathbf{Y}_t = \overline{\mathbf{Y}}_t - \mathbf{H}_t\mathbf{X}_{t,0} = \mathbf{G}_t\mathbf{W}_t \tag{7.97}$$

The system model for $K = 2$ is illustrated in Figure 7.12. Only a small number K of signal sources interfere in a colored-noise manner. Calling $\mathbf{G}_{t,k} \in \mathbb{C}^{N\times n_k}$ the channel from interferer $k \in \{1, \ldots, K\}$, equipped with n_k antennas, to the receiver and $\mathbf{X}_{t,k} \in \mathbb{C}^{n_k\times M}$ the concatenated transmit signals from interferer k, the received signal $\overline{\mathbf{Y}}_t$ can be modeled as

$$\overline{\mathbf{Y}}_t = \mathbf{H}_t\mathbf{X}_{t,0} + \sum_{k=1}^{K} \mathbf{G}_{t,k}\mathbf{X}_{t,k} + \sigma\mathbf{W}_t^{\star} \tag{7.98}$$

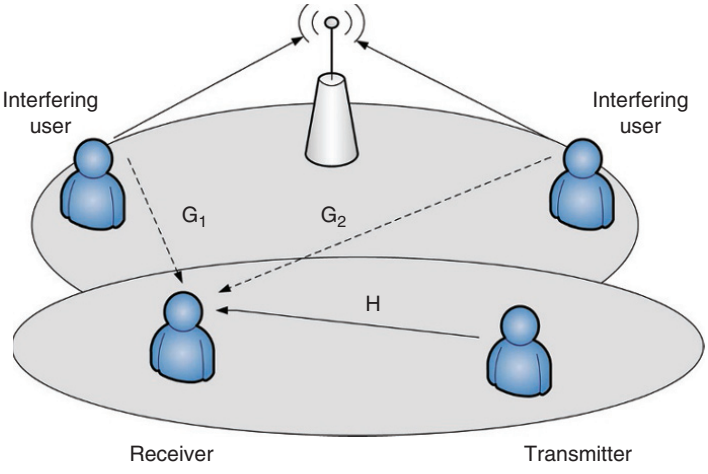


Figure 7.12 System with $K = 2$ interering users.

where $\sigma \mathbf{W}_t^\star \in \mathbb{C}^{N\times M}$ is the concatenated additional white Gaussian noise with variance $\sigma^2 > 0$. In this case, we find that denoting $n = n_1 + \cdots + n_K + N$ and

$$\mathbf{G}_t = \left[\mathbf{G}_{t,1}, \cdots, \mathbf{G}_{t,K}, \sigma \mathbf{I}_N\right], \quad \mathbf{W}_t = \left[\mathbf{X}_{t,1}^T, \cdots, \mathbf{X}_{t,K}^T, \mathbf{W}_t^{\star T}\right]^T$$

we fall back on the aforementioned standard MIMO model (7.97).

The statistical properties of the random variables $\mathbf{X}_{t,0}$ are as follows.

*Assumption **A1**:* For a given t where $1 \le t \le T$, the entries of the matrices $\mathbf{X}_{t,0}$ and $\mathbf{W}_t$ are i.i.d. random variables with a standard complex Gaussian distribution.

The objective for the receiver is to evaluate the average (per antenna) mutual information that can be achieved during the $T \ge 1$ slots. In particular, for $T = 1$, the expression is that of the instantaneous mutual information which allows for an estimation of the rate performance of the current channel. If is T large instead, this provides an approximation of the long-term ergodic mutual information. Under Assumption **A1**, the average mutual information is given by

$$\mathcal{I} = \frac{1}{NT} \sum_{t=1}^{T} \left[\log\det\left(\mathbf{H}_t\mathbf{H}_t^H + \mathbf{G}_t\mathbf{G}_t^H\right) - \log\det\left(\mathbf{G}_t\mathbf{G}_t^H\right)\right] \tag{7.99}$$

Using $\log\det(\cdot) = \mathrm{Tr}\log(\cdot)$, (7.99) is rewritten as

$$\mathcal{I} = \frac{1}{NT} \sum_{t=1}^{T} \left[\mathrm{Tr}\log\left(\mathbf{H}_t\mathbf{H}_t^H + \mathbf{G}_t\mathbf{G}_t^H\right) - \mathrm{Tr}\log\left(\mathbf{G}_t\mathbf{G}_t^H\right)\right] \tag{7.100}$$

The motivation here is to address the problem of estimating $\mathcal{I}$ based on T successive observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_T$ assuming perfect knowledge of, but $\mathbf{H}_1, \ldots, \mathbf{H}_T$, but unknown $\mathbf{G}_t$ for all t.

(7.100) has a form

$$\mathrm{Tr} f\left(\mathbf{A}\mathbf{A}^H\right) = \sum_{i=1}^{n} f\left(\lambda_i\right), \qquad \lambda_i \text{ eigenvalues of } \mathbf{A}\mathbf{A}^H \tag{7.101}$$

with $f(x) = \log(x), x \in \mathbb{R}, x > 0$ where $\mathbf{A} = \frac{1}{\sqrt{n}}\mathbf{\Sigma}_n^{1/2}\mathbf{Z}_n$. Here $\mathbf{\Sigma}_n$ is a non-negative definite Hermitian matrix and $\mathbf{Z}_n$ is a random matrix with i.i.d. real or complex standardized entries. When f is an analytic function, and both dimensions of matrix go to infinity at the same pace, the fluctuations of the linear statistics of (7.101) are shown to be Gaussian. See Example 3.6.3 and Section 3.7.

If the number M of available observations during the sensing period in each slot is very large compared to the channel vector N, a natural estimator, called the standard empirical (SE) estimator, is defined as

$$\hat{\mathcal{I}}_{SE} = \frac{1}{NT} \sum_{t=1}^{T} \left[\log\det\left(\mathbf{H}_t\mathbf{H}_t^H + \frac{1}{M}\mathbf{Y}_t\mathbf{Y}_t^H\right) - \log\det\left(\frac{1}{M}\mathbf{Y}_t\mathbf{Y}_t^H\right)\right] \tag{7.102}$$

When N is fixed, using the law of large numbers and of the continuous mapping theorem, we have as $M \to \infty$

$$\hat{\mathcal{I}}_{SE} - \mathcal{I} \xrightarrow{\text{almost surely}} 0 \tag{7.103}$$

The assumption of $M >> N$ may be feasible in practical settings where sensing needs to be performed fast, particularly under fast-fading conditions. In this case, the standard

empirical estimator is asymptotically biased in the large M, N, regime, hence it is not consistent, and (7.103) will not be valid any more.

Assumptions A2: $M, N, n, n_0 \to +\infty$, and

$$0 < \liminf_{N,n\to\infty} \frac{N}{n} \leqslant \limsup_{N,n\to\infty} \frac{N}{n} < +\infty, \quad 1 < \liminf_{M,N\to\infty} \frac{M}{N} \leqslant \limsup_{M,N\to\infty} \frac{M}{N} < +\infty$$

$$0 < \liminf_{N,n_0\to\infty} \frac{n_0}{N} \leqslant \limsup_{N,n_0\to\infty} \frac{n_0}{N} < +\infty$$

The constraints over N, n and n_0 simply says that these quantities are of the same order. The lower bound for M/N says that M is larger than N, although of the same order.

In the remainder of this section, we refer to Assumption **A2** as the convergence mode $M, N, n \to \infty$.

The channel matrices need be bounded in spectral norm, as $M, N, n \to \infty$.

Assumptions A3: Let $N = N(n)$ a sequence of integers indexed by n. For each $t \in \{1, \ldots, T\}$, consider the family of $N \times n$ matrices $\mathbf{G}_t$. Then, we have the following.

- The spectral norms of $\mathbf{G}_t$ are uniformly bounded in the sense that

$$\sup_{1\leqslant t\leqslant T} \sup_{N,n} \left\|\mathbf{G}_t\right\| < \infty$$

- For $t \in \{1, \ldots, T\}$, the smallest eigenvalue of $\mathbf{G}_t \mathbf{G}_t^H$ denoted by $\lambda_N\left(\mathbf{G}_t \mathbf{G}_t^H\right)$ is uniformly bounded away from zero, i.e., there exists $\sigma^2 > 0$ such that

$$\inf_{1\leqslant t\leqslant T} \inf_{N,n} \left\|\mathbf{G}_t \mathbf{G}_t^H\right\| \geqslant \sigma^2 > 0$$

Assumptions A4: Let $N = N(n)$ a sequence of integers indexed by n_0. For each $t \in \{1, \ldots, T\}$, consider the family of $N \times n_0$ matrices $\mathbf{H}_t$. Then, the spectral norms of $\mathbf{H}_t$ are uniformly bounded in the sense that

$$\sup_{1\leqslant t\leqslant T} \sup_{N,n_0} \left\|\mathbf{H}_t\right\| < \infty$$

Assumptions A5: The family of matrices $(\mathbf{H}_t)$ satisfies additionally the following assumptions.

- Consider the rank of $\mathbf{H}_t$. Then

$$0 < \liminf_{N,n_0\to\infty} \frac{\operatorname{rank}\left(\mathbf{H}_t\right)}{N} \leqslant \limsup_{N,n_0\to\infty} \frac{\operatorname{rank}\left(\mathbf{H}_t\right)}{N} < 1$$

- The smallest eigenvalue of $\mathbf{H}_t \mathbf{H}_t^H$ is uniformly bounded away from zero, i.e., there exists $\kappa > 0$ such that

$$\inf_{1\leqslant t\leqslant T} \inf_{N,n_0} \left\|\lambda_i\left(\mathbf{H}_t \mathbf{H}_t^H\right) \left|\lambda_i\left(\mathbf{H}_t \mathbf{H}_t^H\right) > 0\right\| \geqslant \kappa > 0$$

The main result of this section is presented here.

Theorem 7.8.3 (G-estimator for the average mutual information) Assume that **A1**–**A5** hold and define the quantity

$$\hat{I}_G = \frac{1}{NT} \sum_{t=1}^{T} \log\det\left(\mathbf{I}_N + \hat{\gamma}_{N,t} \mathbf{H}_t \mathbf{H}_t^H \left(\frac{1}{M} \mathbf{Y}_t \mathbf{Y}_t^H\right)^{-1}\right)$$

$$+ \frac{1}{T} \sum_{t=1}^{T} \frac{(M-N)}{N} \left[\log\left(\frac{M}{M-N} \hat{\gamma}_{N,t}\right) + 1\right] - \frac{M}{N} \hat{\gamma}_{N,t}$$

where $\hat{y}_{N,t}$ is the unique real positive solution of

$$\hat{y}_{N,t} = \frac{\hat{y}_{N,t}}{M} \operatorname{Tr}\left[\mathbf{H}_t \mathbf{H}_t^H \left(\hat{y}_{N,t} \mathbf{H}_t \mathbf{H}_t^H + \frac{1}{M} \mathbf{Y}_t \mathbf{Y}_t^H\right)^{-1}\right] + \frac{M-N}{M}$$

Then

$$\hat{\mathcal{I}}_G - \mathcal{I} \xrightarrow[M,N,n\to\infty]{\text{Almost Surely}} 0.$$

Theorem 7.8.4 (central limit theorem) Assume that **A1**–**A5** hold true. Then

$$\frac{N}{\sqrt{\theta_N}} \left(\hat{\mathcal{I}}_G - \mathcal{I}\right) \xrightarrow[N\to\infty]{\text{Distribution}} \mathcal{N}(0,1)$$

where θ_N is given by

$$\theta_N = \frac{1}{T^2} \sum_{t=1}^{T} 2 \log\left(M \hat{y}_{N,t}\right) - \log\left[(M-N)\left(M - \operatorname{Tr}\left[\left(\mathbf{I}_N + \mathbf{H}_t \mathbf{H}_t^H \left(\mathbf{G}_t \mathbf{G}_t^H\right)^{-1}\right)^{-2}\right]\right)\right]$$

which is a well-defined quantity which satisfies

$$0 < \liminf_{M,N,n\to\infty} \theta_N \leqslant \limsup_{M,N,n\to\infty} \theta_N < +\infty$$

Example 7.8.5 (Massive MIMO) This section is motivated by a massive MIMO system, whose receiver is equipped with a large number N of antenna, such as $N = 200 - 1,000$, and whose transmitter is equipped with n_0 antennas. The receiver is exposed to K interfering signals. Consider the uplink of the system illustrated in Figure 7.12. □

Bibliographical Remarks

In Section 7.3, we follow [384].

Section 7.4, is mainly based on [391]. In Section 7.4.5, we follow [398].

We follow [425] in Section 7.5.

El Karoui [178] has proposed a variational and nonparametric approach to this problem based on an appropriate distance function using the Marchenko–Pastur equation (4.5). In another important work [426] the authors propose useing a suitable set of empirical moments, say the first q moments: for $k = 1, \ldots, q$,

$$\hat{\alpha}_k = \frac{1}{N} \operatorname{Tr}\left(\mathbf{S}_n^k\right) = \frac{1}{N} \sum_{i=1}^{N} \lambda_i^k \left(\mathbf{S}_n\right)$$

where λ_i are the eigenvalues of $\mathbf{S}_n$ (assuming $N \leq n$).

In [53], a modification of the procedure in [426] is proposed to obtain a direct moments estimator based on the sample moments $(\hat{\alpha}_k)$. Compared to El Karoui [178] and Rao *et al.* [426], this moment estimator is simpler and much easier to implement. Moreover, the convergence rate of this estimator (asymptotic normality) is also established. Chen *et al.* [427] have also analyzed the underlying order selection problem

and proposed a solution based on the crossvalidation principle. We have followed Li *et al.* (2013) [220] and [228] in Section 4.3. The new approach of Li *et al*. (2013) [220] can be viewed as a synthesis of the optimization approach in El Karoui [178] and the parametric setup in [53]. On one hand, the authors of [220] adopt the optimization approach and prove that it is in general preferable to the moment approaches. On the other hand, using a generic parametric approach for discrete population spectral densities as well as continuous population spectral densities, they are able to avoid the implementation difficulties in El Karoui [178]. Another important advantage of [220] is that the optimization problem has been moved from the complex plan to the *real line* by considering a characteristic equation (the Marchenko–Pastur equation) on the real line. The optimization procedure obtained is then much simpler than the original one in [178].

Consider the general matrix

$$\mathbf{S_n} = \frac{1}{n}\mathbf{T}_n^{1/2}\mathbf{X}_n\mathbf{X}_n^H\mathbf{T}_n^{1/2}$$

where $\mathbf{X}_n = \left(x_{ij}\right)$ is a $p \times p$ matrix consisting of *independent* complex entries with mean zero and variance one, $\mathbf{T}_n$ is a $p \times p$ nonrandom positive definite Hermitian matrix with spectral norm uniformly bounded in p. Note that the entries of $\mathbf{X}_n$are not necessarily i.i.d. In [428], assuming the eighth moment, the authors find that the rate of the expected empirical spectral distribution of $\mathbf{S}_n$ converging to its limit spectral distribution is $O(1/\sqrt{n})$. Moreover, under the same assumption, we prove that for any $\varepsilon > 0$, the rates of the convergence of the empirical spectral distribution of $\mathbf{S}_n$ in probability and the almost sure convergence are $O\left(1/n^{2/5}\right)$ and $O\left(1/n^{2/5+\varepsilon}\right)$ respectively.

In Section 7.6, we follow [402] and [416] for the development. We take material from [372] for Example 8.9.2.

In Section 7.7, we take material from [405, 406, 411, 416]. References [411] and [415] can be viewed as two companion papers. Two papers from Ledoit and Wolf are [429, 430]. High-dimensional covariance matrix estimation with missing observations is treated in [431].

The material in Section 7.8 is taken from [418–420]. Example 7.8.1 is taken from [419, 420]. Example 7.8.2 is taken from [421]. The model follows [422] closely. Reference [421] is relevant in this context.

In Section 7.2, we draw some material from [381, 432].

8

Matrix Hypothesis Testing using Large Random Matrices

This chapter can be viewed as the first application of big data. This chapter is relevant to all the three applications we have in mind: smart grid, communications, and sensing. We assume that massive datasets are at our disposal for data processing. The problem may be conveniently formulated in terms of a matrix-valued hypothesis testing problem. Here we emphasize the use of large random matrix $\mathbf{X}$ as an elementary mathematical object to study. We view $\mathbf{X}$ as a whole and study the matrix-valued function $f(\mathbf{X})$. See Section 8.2 for an example.

A new line of research in power system security has focused on cyber intrusion related to intelligent electronic devices, such as remote terminal units, phasor measurement units, and meters. In this chapter, we are motivated by big data aspects of the power system grid. In particular, our goal is to model the large datasets using large dimensional random matrices. The fundamentals of large random matrices were treated in Chapter 3. The high dimensionality of the new formulation will unveil some unique features of the problems. We, therefore, investigate the cyber security of the smart grid from the viewpoint of anomaly detection, emphasizing the high dimensionality in the framework of large random matrices. The application of these recent results (in the random matrix theory literature) to the context of smart grid appears novel, as far as the author is aware.

The Anomaly Detection at Multiple Scales program at the Defense Advanced Projects Research Agency [42] creates, adapts and applies technology to anomaly characterization and detection in massive data sets. Anomalies in data cue the collection of additional, actionable information in a wide variety of real-world contexts. The initial application domain is insider threat detection in which malevolent (or possibly inadvertent) actions by a trusted individual are detected against a background of everyday network activity.

Let the data speak (and only the data). This is a sound principle, provided that there is enough data to trust the data.

The infinite-dimensional Hilbert operators are replaced with large, but finite random matrices. By analogy, the massive data sets are naturally represented by large random matrices. Data representation using large random matrices seems novel in the context of big data. This representation facilitates dimension reduction and the study of scalability, through the functions of their eigenvalues.

We view these metrics as (possibly nonlinear) functions of random matrices. Trace function is linear. It is not obvious how to choose these functions for good performance.

Smart Grid using Big Data Analytics: A Random Matrix Theory Approach, First Edition.
Robert C. Qiu and Paul Antonik.

The analysis of these metrics requires advanced tools, such as the nonasymptotic theory of random matrix. Concentration of measure is the foundation for this theory.

We demand a theory of random matrices, which is valid for arbitrary sizes of matrices. The so-called scalability required for big data (Chapter 1.1) justifies the significance of this nonasymptotic random matrix theory.

8.1 Motivating Examples

A large number of data samples can be summarized in the form of a large-dimensional random matrix.

Example 8.1.1 (massive MIMO) For the large antenna array of $N = 1,000$ antennas at the base station, we assume K mobile users. Often we use multicarrer OFDM systems with say $M = 128$ subcarriers. We collect together all the data for the k-th user into an random vector $\mathbf{x}_k$

$$\mathbf{x}_k = \left[X_{1k}, ..., X_{(MN),k}\right]^T \in \mathbb{C}^{MN}$$

and then form a random matrix $\mathbf{X}$

$$\mathbf{X} = \left[\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_K\right] \in \mathbb{C}^{MN \times K}$$ □

Example 8.1.2 (time series) For the above massive MIMO example, for each epoch t, we associate the data matrix with $\mathbf{X}_t \in \mathbb{C}^{MN \times K}$ for $t = 1, \dots, T$. We obtain a sequence of large random matrices. If we put all the data together, a three-dimensional array (or tensor) can be used to summarize the data. For the k-th user, we can study

$$\mathbf{X}_k = \left[\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T\right] \in \mathbb{C}^{MN \times T}$$

for every $k = 1, \dots, K$. It is interesting to study the setting that T and MN are comparable. □

Example 8.1.3 (matrix-valued distributions for dependence structure of the data) Multivariate normal distribution plays a central role in the theory of multivariate statistical analysis. Even when the original data is not multivariate normal, due to the central limit theorem, sampling distributions of certain statistics can be approximated by normal distribution. When the central limit theorem is not valid, the next best thing to do is the concentration inequalities of certain matrix functions that describe the concentrations of measure phenomenon around their expectations [40, p. 146].

The independent multivariate observations are often written in terms of a matrix, which is known as a sample observation matrix. In such a matrix, when sampling from multivariate normal distribution, the columns are distributed *independently* as multivariate normal with common mean vector and covariance matrix. The assumption of independence of multivariate observations is not met in multivariate time series, stochastic processes and repeated measurements on multivariate variables. In these cases, the matrix of observations lead to the introduction of the matrix variate normal distribution.

Random matrices can be used to describe repeated measurements on multivariate variables. The assumption of the independence of the observations is often not feasible. When analyzing data sets lie these, the matrix variate elliptically contoured distributions can be used to describe the *dependence structure* of the data.

Matrix variate elliptically contoured distributions represent an extension of the concept of elliptical distributions from the vector to the matrix case [217,433]. The fact that the distributions in this class possess certain properties, similar to those of the Gaussian distribution, makes them especially useful. For example, many testing procedures developed for the Gaussian theory to test various hypotheses can be used for this class of distributions too.

For the i-th time series $\mathbf{x}_i \in \mathbb{C}^{1\times T}$, we can repeat N measurements to obtain the random matrix $\mathbf{X}_i \in \mathbb{C}^{N\times T}$. In other words, we extend the random vector to the random matrix

$$\mathbf{x}_i \in \mathbb{C}^{1\times T} \to \mathbf{X}_i \in \mathbb{C}^{N\times T}, \quad i = 1, \dots, n$$

Although straightforward, the extension plays a central role in modeling the massive amount of data. We are interested in the cases where N, T and n are large, say in the order of 1000. A total of 10^9 or one billion data points can be easily handled by today's computing capability.

This model has been used for stock markets when we assume the matrix of returns follows a matrix of elliptically contoured distributions. This family turns out to be very suitable to describe stock returns because the returns are *neither assumed to independent or to be normally distributed*. We See Section 8.12 for details.

The applications include:

- time-varying complex network;
- large-scale intrusion detection and anomaly detection;
- preventing large-scale denial of service. □

8.2 Hypothesis Test of Two Alternative Random Matrices

We now consider the so-called matrix hypothesis testing problem:

$$\begin{aligned} \mathcal{H}_0 &: \mathbf{X} \\ \mathcal{H}_1 &: \mathbf{Y} = \sqrt{\mathrm{SNR}} \cdot \mathbf{H} + \mathbf{X} \end{aligned} \tag{8.1}$$

where SNR represents the signal-to-noise ratio, and $\mathbf{H}$ and $\mathbf{X}$ are two non-Hermitian random matrices of $m \times n$. We further assume that $\mathbf{H}$ is independent of $\mathbf{X}$. The problem of (1.12) is equivalent to

$$\begin{aligned} \mathcal{H}_0 &: \mathbf{X}\mathbf{X}^H \\ \mathcal{H}_1 &: \mathbf{Y}\mathbf{Y}^H = \mathrm{SNR} \cdot \mathbf{H}\mathbf{H}^H + \mathbf{X}\mathbf{X}^H + \sqrt{\mathrm{SNR}}\left(\mathbf{H}\mathbf{X}^H + \mathbf{X}\mathbf{H}^H\right) \end{aligned} \tag{8.2}$$

where $\mathbf{H}\mathbf{H}^H, \mathbf{X}\mathbf{X}^H, \mathbf{Y}\mathbf{Y}^H$ are positive semidefinite random matrices. A matrix $\mathbf{A}$ of $m \times n$ is said to be positive semidefinite if all the eigenvalues of $\mathbf{A}$ are non-negative, i.e., $\lambda_i(\mathbf{A}) \geq 0, i = 1, \dots, \min(m, n)$. All matrices in (1.12) are non-Hermitian, while this is not true in (1.13). In extremely low SNR, the cross terms $\mathbf{H}\mathbf{X}^H$ and $\mathbf{X}\mathbf{H}^H$ in (1.13)

are non-Hermitian random matrices and may be dominant in the performance of the detection. Most algorithms in the past focused on the formulation of (1.13); Perhaps in the future we may pay more attention to (1.12) using the non-Hermitian random matrix theory.

We need a statistic metric for hypothesis detection: decide on the hypothesis of random matrix $\mathbf{X}\mathbf{X}^H$ ($\mathcal{H}_0$) or random matrix$\mathbf{Y}\mathbf{Y}^H$ ($\mathcal{H}_0$). Scalar metrics are more desirable than vectors and matrices. Natural candidates for these scalar metrics are: (i) individual eigenvalues $\lambda_i(\mathbf{Y}\mathbf{Y}^H), i = 1, \ldots, \min(m, n)$; (ii) the trace $\mathrm{Tr}\left(\mathbf{X}^H\mathbf{X}\right)$ or $\mathrm{Tr}\left(\mathbf{Y}^H\mathbf{Y}\right)$. We view these metrics as (possibly nonlinear) functions of random matrices. Trace function is linear. It is not obvious how to choose these functions for good performance.

The analysis of these metrics requires advanced tools, such as nonasymptotic theory of random matrix. Concentration of measure is the foundation for this theory.

8.3 Eigenvalue Bounds for Expectation and Variance

The aim of this section is to provide sharp nonasymptotic bounds for the variance of individual eigenvalues of sample covariance matrices, following [434, 435]. Eigenvalues are treated as scalar-valued functions of a random matrix $\mathbf{X}$. This topic has been studied extensively using concentration inequalities in [40]. Literature on this topic has been comprehensively surveyed in [40].

Random covariance matrices, or Wishart matrices, were introduced by the statistician Wishart in 1928 to model tables of random data in multivariate statistics. The spectral properties of these matrices are critical to statistical tests and for principal component analysis. Eigenvalues were studied asymptotically both at the global and local regimes, considering the global behavior of the spectrum, the behavior of extreme eigenvalues or the spacings between eigenvalues in the bulk of the spectrum. In the Gaussian case, the eigenvalue joint distribution is explicitly known, allowing for a complete study of the asymptotic spectral properties (see [35, 163, 436]. One of the main goals of random matrix theory over the past decades was to extend these results to non-Gaussian covariance matrices.

We demand a theory of random matrices, which is valid for arbitrary sizes of matrices. The so-called scalability required for big data (Section 1.1) justifies the significance of this nonasymptotic theory of random matrices.

Let $\mathbf{Z}$ be a $N \times n$ (real or complex) data matrix, with $N \geq n$, such that its entries are independent, centered and have variance 1. The sample covariance matrix is defined as $\mathbf{S} = \frac{1}{N}\mathbf{Z}^H\mathbf{Z}$.

The hypothesis $\mathcal{H}_0$ of (1.13) is the case when the entries of $\mathbf{X}$ are Gaussian. Recall that our goal in Section 8.2 is to evaluate the performance of the metrics for hypothesis detection: these metrics are treated as functions of random matrices. For any function $f(x)$, we can replace the scalar x with a matrix $\mathbf{A}$ of $n \times n$. The matrix function $f(\mathbf{A})$ is described by the function of eigenvalues $f\left(\lambda_1, \ldots, \lambda_n\right)$. So we deal with scalar random variables $\lambda_1, \ldots, \lambda_n$, which are dependent on each other.

For a scalar random variable X, the expectation $\mu = \mathbb{E}X$ and the variance $\sigma^2 = \mathrm{Var}\, X$ are of fundamental interest in the context of hypothesis tests. In practice, we must estimate them. For n samples or realizations of X, $x_1, \ldots, x_n$, the expectation is estimated

by $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and the variance is estimated by $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i\right)^2$. The classical limit theorems of probability [437–440] deal with a sequence of independent random variables. Two of the most important propositions in probability theory are the law of large numbers and the central limit theory [441].

Let us study an example to motivate this whole section.

Example 8.3.1 (matrix hypothesis test at multiple scales) Let us revisit the motivating problem in Section 8.2. We illustrate how eigenvalue inequalities are used to design algorithms.

It is desirable to have a metric for hypothesis tests as a function of matrix size n. When we are given a large data set, we start with a sample covariance random matrix. For convenience, we reproduce (1.13) as follows

$$\begin{aligned} \mathcal{H}_0 &: \mathbf{X}\mathbf{X}^H \\ \mathcal{H}_1 &: \mathbf{Y}\mathbf{Y}^H = \mathrm{SNR}\cdot\mathbf{H}\mathbf{H}^H + \mathbf{X}\mathbf{X}^H + \sqrt{\mathrm{SNR}}\left(\mathbf{H}\mathbf{X}^H + \mathbf{X}\mathbf{H}^H\right) \end{aligned} \tag{8.3}$$

where $\mathbf{H}\mathbf{H}^H, \mathbf{X}\mathbf{X}^H, \mathbf{Y}\mathbf{Y}^H$ are positive semidefinite random matrices. For sample covariance random matrix $\mathbf{X}\mathbf{X}^H$ of $n \times n$, let λ_i be the sorted eigenvalues and γ_i be their theoretical predictions. There exists a constant C such that

$$\sum_{i=1}^{n} \mathbb{E}\left[\left(\lambda_i - \gamma_i\right)^2\right] \leqslant C\frac{\log n}{n} \tag{8.4}$$

which is identical to (8.16) below. Our algorithm is defined as follows

$$\begin{aligned} \mathcal{H}_0 &: \sum_{i=1}^{n} \mathbb{E}\left[\left(\lambda_i - \gamma_i\right)^2\right] \leqslant \Lambda \\ \mathcal{H}_1 &: \sum_{i=1}^{n} \mathbb{E}\left[\left(\lambda_i - \gamma_i\right)^2\right] > \Lambda \end{aligned} \tag{8.5}$$

where the decision threshold Λ is defined as

$$\Lambda = C\frac{\log n}{n}$$

Using (8.12) below, we have

$$\mathrm{Var}\left(\lambda_i\right) \leqslant C\frac{\log n}{n^2} \tag{8.6}$$

Similarly, we can design a hypothesis test algorithm as follows

$$\begin{aligned} \mathcal{H}_0 &: \mathrm{Var}\left(\lambda_i\right) \leqslant \Lambda \\ \mathcal{H}_1 &: \mathrm{Var}\left(\lambda_i\right) > \Lambda \end{aligned}$$

where

$$\Lambda = C\frac{\log n}{n^2}$$

We conclude here that the decision threshold is an explicit function of the size n of a large data set that is represented as a random data matrix $\mathbf{X}$. When we raise n to infinity,

we obtain the asymptotic regime. In practice, however, the central interest is in the scaling speed as a function n. it is often studied in nonasymptotic theory of random matrix, see for example [40, 442].

For hypothesis $\mathcal{H}_1$, the problem at hand is related to the outliers of $\frac{1}{\sqrt{n}}\mathbf{Z} + \mathbf{A}$, where $\mathbf{Z}$ is an i.i.d. random matrix of $n \times n$ and $\mathbf{A}$ is a low-rank perturbation. See [333] for the outlier problem. □

8.3.1 Theoretical Locations of Eigenvalues

Let $\mathbf{X}$ be a $m \times n$ (real or complex) matrix, with $m \geq n$, such that its entries are independent, centered, and have variance 1. Then $\mathbf{S} = \frac{1}{m}\mathbf{X}^H\mathbf{X}$ is a sample covariance matrix. An important example is the case when the entries of $\mathbf{X}$ are Gaussian. Then $\mathbf{S}$ belongs to the so-called Laguerre unitary ensemble (LUE) if the entries of $\mathbf{X}$ are complex and to the Laguerre orthogonal ensemble (LOE) if they are real. $\mathbf{S}$ is Hermitian (or real symmetric) and therefore has n real eigenvalues. As $m \geq n$, none of these eigenvalues is trivial. These eigenvalues are non-negative and will be denoted by $0 \leqslant \lambda_1 \leqslant \cdots \leqslant \lambda_n$.

Among universality results, the classical Marchenko–Pastur theorem states that if $\frac{m}{n} \to y \geqslant 1$ when n goes to infinity, the empirical spectral distribution

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i} \tag{8.7}$$

almost certainly converges to a deterministic measure $\mu_{MP}(x)$, called the Marchenko–Pastur distribution of parameter y. $\hat{\mu}$ is a random probability measure. The measure $\mu_{MP}(x)$ is compactly supported and is absolutely continuous with respect to Lebesgue measure, with density

$$d\mu_{MP}(x) = \frac{1}{2\pi x}\sqrt{(x-a)(b-x)}\mathbf{1}_{[a,b]}(x)\,dx$$

with $a = \left(1-\sqrt{y}\right)^2, b = \left(1+\sqrt{y}\right)^2$. We denote by $\mu_{m,n}$ the approximate Marchenko–Pastur density

$$\rho_{m,n}(x) = \frac{1}{2\pi x}\sqrt{\left(x-a_{m,n}\right)\left(b_{m,n}-x\right)}\mathbf{1}_{[a_{m,n},b_{m,n}]}(x)$$

with $a_{m,n} = \left(1-\sqrt{\frac{m}{n}}\right)^2, b_{m,n} = \left(1+\sqrt{\frac{m}{n}}\right)^2$. The behavior of individual eigenvalues was more difficult to achieve.

The following law of large numbers is obtained. For all $\eta > 0$, and all $\eta n \leqslant i \leqslant (1-\eta)n$, i.e., the eigenvalues in the bulk of the spectrum,

$$\lambda_i - \gamma_i \xrightarrow[n\to\infty]{} 0, \text{ almost certainly,}$$

where the theoretical location $\gamma_i \in \left[a_{m,n}, b_{m,n}\right]$ of the i-th eigenvalue λ_i is defined by

$$\frac{i}{n} = \int_{a_{m,n}}^{\gamma_i} \rho_{m,n}(x)dx \tag{8.8}$$

8.3.2 Wasserstein Distance

Assume we have access to the big data in $\mathbb{R}^n$, whose law is supposed to be the approximation of a probability measure μ of interest. $\hat{\mu}$ is defined in

$$\hat{\mu}^n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i} \tag{8.9}$$

which is a random probability measure. Suppose that the empirical measure associated with our big data system is a good approximation of μ, with very high probability. More precisely,

$$\mathbb{P}\left(W_p\left(\hat{\mu}^n, \mu\right) \geqslant \varepsilon\right) \leqslant \tau_p(n, \varepsilon) \tag{8.10}$$

where $\tau_p(n, \varepsilon)$ is a known function of n and ε, and $\mathbb{P}$ is the probability measure on the probability space.

8.3.3 Sample Covariance Matrices—Entries with Exponential Decay

For simplicity, we assume that $y > 1$. More precisely, we assume that $1 < \alpha \leq m/n \leq \beta$ where α, β are fixed constants. Assume furthermore that $\mathbf{S}$ is a complex covariance matrix whose entries have an exponential decay and have the same first four moments as those of an LUE (alternatively Laguerre Orthogonal Ensemble) matrix. This condition is called condition (C0). Matrices which are considered in this subsection are sample covariance matrices $\mathbf{S}$ satisfying condition (C0). We say that $\mathbf{S} = \frac{1}{m}\mathbf{X}^H\mathbf{X}$ satisfies condition (C0) if the entries X_{ij} of $\mathbf{X}$ are independent and have an exponential decay: there are positive constants C_1 and C_2 such that

$$\forall i \in \{1, \ldots, n\}, \forall j \in \{1, \ldots, m\}, \quad \mathbb{P}\left(\left|X_{ij}\right| \geqslant t^{C_1}\right) \leqslant e^{-t} \tag{8.11}$$

for all $t \geq C_2$. For variance, compare the results below with (3.57).

In the bulk of the spectrum. Let $\eta \in (0, 1/2)$. There exists a constant $C > 0$ (depending on η, α, β) such that for every sample covariance matrix $\mathbf{S}$, for every $\eta n \leqslant i \leqslant (1-\eta)n$

$$\mathrm{Var}\left(\lambda_i\right) \leqslant C\frac{\log n}{n^2} \tag{8.12}$$

Between the bulk and the edge of the spectrum. There exists a constant $\kappa > 0$ (depending on α, β) such that the following holds. For all $K \geq \kappa$, and $\eta \in (0, 1/2]$, these is a constant $C > 0$ such that for every sample covariance matrix $\mathbf{S}$, for every $(1-\eta)\, n \leqslant i \leqslant n - K\log n$

$$\mathrm{Var}\left(\lambda_i\right) \leqslant C\frac{\log(n-i)}{n^{4/3}(n-i)^{2/3}} \tag{8.13}$$

where the constant C depends on K, η, α, β.

At the edge of the spectrum. There exists a constant $C > 0$ (depending on α, β) such that, for every sample covariance matrix $\mathbf{S}$

$$\mathrm{Var}\left(\lambda_i\right) \leqslant C\frac{1}{n^{4/3}} \tag{8.14}$$

Rate of convergence towards the Marchenko–Pastur distribution, Let γ_i be defined in (8.8). The bounds on $\mathbb{E}\left[\left(\lambda_i - \gamma_i\right)^2\right]$ lead to a bound on the rate of convergence of the empirical spectral measure $\mathcal{L}$ towards the Marchenko–Pastur distribution in terms of 2-Wasserstein distance. $W_2^2(\mathcal{L}, \mu)$ is a random variable defined by

$$W_2(\mathcal{L}, \mu) = \inf_{\pi}\left(\int_{\mathbb{R}^2} |x-y|^2 d\pi\left(x, y\right)\right)^{1/2}$$

where the infimum is taken over all probability measures π on $\mathbb{R}^2$ such that its first marginal is $\mathcal{L}$ and its second marginal is μ. To achieve the expected bound, we rely on another expression of W_2 in terms of distribution functions, namely

$$W_2^2(\mathcal{L},\mu) = \int_0^1 \left(F^{-1}(x) - G^{-1}(x)\right)^2 dx$$

where $F^{-1}(x)$ (respectively $G^{-1}(x)$) is the generalized inverse of the distribution function $F(x)$ (respectively $G(x)$) of $\mathcal{L}$ (respectively μ) [443]. These functions depend on the matrix size m, n.

There exists a constant $C > 0$ depending only on β such that for all $1 \leq \frac{m}{n} \leq \beta$

$$W_2^2(\mathcal{L},\mu) \leqslant \frac{2}{n}\sum_{i=1}^{n}\left(\lambda_i - \gamma_i\right)^2 + \frac{C}{n^2} \tag{8.15}$$

The intuition to study the sum of (possible dependent) random variables $\left(\lambda_i - \gamma_i\right)^2$ is satisfactory. The average $\frac{1}{n}\sum_{i=1}^{n}\left(\lambda_i - \gamma_i\right)^2$ will greatly improve the accuracy of the estimate of individual term $\left(\lambda_i - \gamma_i\right)^2$ when n turns large.

Let $1 < \alpha < \beta$. Then exists a constant $C > 0$ depending only on α and β such that, for all m and n such that $1 < \alpha \leq \frac{m}{n} \leq \beta$,

$$\sum_{i=1}^{n}\mathbb{E}\left[\left(\lambda_i - \gamma_i\right)^2\right] \leqslant C\frac{\log n}{n} \tag{8.16}$$

Therefore

$$\mathbb{E}\left[W_2^2(\mathcal{L},\mu)\right] \leqslant C\frac{\log n}{n^2} \tag{8.17}$$

The convergence rate of $\log n/n^2$ is very impressive when n turns large, say 1000.

Comparing (8.17) with (3.56) for the expectation, we find that the rate of the expectation of moments decays with $O(1/n)$, in contrast with $\frac{\log n}{n^2}$ of $\mathbb{E}\left[W_2^2(\mathcal{L},\mu)\right]$.

8.3.4 Gaussian Covariance Matrices

This subsection is concerned with Gaussian covariance matrices. The results and techniques used here rely heavily on the Gaussian structure, in particular on the determinantal properties of the eigenvalues. Let γ_i be defined in (8.8).

Inside the bulk of the spectrum. Let $\eta \in (0, 1/2]$ and $1 < \alpha < \beta$. Let $\mathbf{S}$ be an LUE matrix. There exists a constant $C > 0$ (depending only on η, α, β) such that for all $\alpha \leqslant \frac{m}{n} \leqslant \beta$ and for every $\eta n \leqslant i \leqslant (1-\eta)n$

$$\mathbb{E}\left[\left|\lambda_i - \gamma_i\right|^2\right] \leqslant C\frac{\log n}{n^2} \tag{8.18}$$

In particular

$$\mathrm{Var}\left(\lambda_i\right) \leqslant C\frac{\log n}{n^2} \tag{8.19}$$

Between the bulk and the edge of the spectrum. There exists a constant $\kappa > 0$ (depending on α, β) such that the following holds. For all $K \geq \kappa$, and $\eta \in (0, 1/2]$

these is a constant $C > 0$ such that for every sample covariance matrix $\mathbf{S}$, for every $(1-\eta)\, n \leqslant i \leqslant n - K \log n$

$$\mathbb{E}\left[\left(\lambda_i - \gamma_i\right)^2\right] \leqslant C \frac{\log(n-i)}{n^{4/3}(n-i)^{2/3}} \tag{8.20}$$

In particular

$$\mathrm{Var}\left(\lambda_i\right) \leqslant C \frac{\log(n-i)}{n^{4/3}(n-i)^{2/3}} \tag{8.21}$$

where the constant C depends on K, η, α, β.

At the edge of the spectrum. Let $\alpha > 1$. There exists a constant $C > 0$ depending only on α such that the following holds. Let $\mathbf{S}$ be an LUE matrix. Denote by λ_{max} the maximal eigenvalue of $\mathbf{S}$. Then, for all $n \in \mathbb{N}$ such that $m > \alpha n$, and for all $0 < \varepsilon \leq 1$,

$$\mathbb{P}\left(\lambda_{\max} \leqslant b_{m,n}(1-t)\right) \leqslant C^2 \exp\left(-\frac{2}{C} n^2 t^3\right) \tag{8.22}$$

and

$$\mathbb{P}\left(\lambda_{\max} \geqslant b_{m,n}(1+t)\right) \leqslant C \exp\left(-\frac{2}{C} n t^{3/2}\right)$$

The large deviation tails are also known. Let $\mathbf{S}$ be an LUE matrix. Then there exists a universal constant $C > 0$ such that for all $n \geq 1$, for all $m \in \mathbb{N}$ such that $m > \alpha n$

$$\mathrm{Var}\left(\lambda_{\max}\right) \leqslant \mathbb{E}\left[\left(\lambda_{\max} - b_{m,n}\right)^2\right] \leqslant C \frac{1}{n^{4/3}} \tag{8.23}$$

Similar results are probably true for the k-th largest eigenvalue (for $k \in \mathbb{N}$ fixed). A left-side deviation inequality for the smallest eigenvalue is established also in the case when $m > \alpha n$

$$\mathbb{P}\left(\lambda_{\min} \leqslant a_{m,n}(1-t)\right) \leqslant C \exp\left(-\frac{2}{C} n t^{3/2}\right) \tag{8.24}$$

for all $0 < t \leq 1$. But no right-side deviation inequality seems to be known for the smallest eigenvalue $\lambda_{\min}$ and therefore we cannot deduce a precise bound on the variance of the smallest eigenvalue.

8.4 Concentration of Empirical Distribution Functions

In this section we are interested in concentration as a function of matrix size n. Our method is to convert a random matrix into a random vector of eigenvalues.

Let $\mathbf{X} = \left(X_1, \ldots, X_n\right)$ be a random vector in $\mathbb{R}_n$ with distribution μ. We study rates of approximation of the average marginal distribution function

$$F(x) = \mathbb{E}F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}\left\{X_i \leqslant x\right\}$$

by the empirical distribution function

$$F_n(x) = \frac{1}{n} \mathrm{card}\left\{i \leqslant n \,:\, X_i \leqslant x\right\}, \quad x \in \mathbb{R}$$

where card(·) denotes the cardinality of the set. We shall measure the distance between F and F_n by means of the (uniform) Kolmogorov metric

$$\left\|F_n - F\right\| = \sup_x \left|F_n(x) - F(x)\right|$$

and as well as by means of the L_1-metric

$$W_1\left(F_n, F\right) = \int_{-\infty}^{\infty} \left|F_n(x) - F(x)\right| dx$$

The latter, also called the Kantorovich–Rubinstein distance, may be interpreted as the minimal cost needed to transport the empirical measure F_n to F with cost function

$$d(x, y) = |x - y|$$

(the price paid to transport the point x to the point y).

The classical example is the case where all X_is are independent and identically distributed (i.i.d.), that is, when μ represents a product measure on $\mathbb{R}^n$ with equal marginals, say, F.

On the other hand, the observations $X_1, \ldots, X_n$ may also be generated by nontrivial functions of *independent* random variables. Of particular importance are random symmetric matrices $\left(\frac{1}{\sqrt{n}}\xi_{ij}\right), 1 \leqslant i, j \leqslant n$, with i.i.d. entries above and on the diagonal. Arranging their eigenvalues $X_1 \leqslant \cdots \leqslant X_n$ in increasing order, we arrive at the spectral empirical measures F_n. In this case, the mean $F = \mathbb{E}F_n$ also depends on n and converges to the semicircle law under appropriate moment assumptions on ξ_{ij}. By studying the spectral empirical measures, we reduce a random matrix problem to a simpler random vector problem.

The study of matrices motivates the study of deviations of F_n from the mean F under general analytical hypotheses on the joint distribution of the observations, such as Poincare or logarithmic Sobolev inequalities. A probability measure μ on $\mathbb{R}^n$ is said to satisfy a Poincare-type or spectral gap inequality with constant σ^2 ($\sigma > 0$) if, for any bounded smooth function g on $\mathbb{R}^n$ with gradient ∇g

$$\mathrm{Var}_\mu\left(g\right) \leqslant \sigma^2 \int |\nabla g|^2 d\mu \tag{8.25}$$

Similarly, μ satisfies a logarithmic Sobolev inequality with constant σ^2 if, for all bounded smooth g

$$\mathrm{Ent}_\mu\left(g^2\right) \leqslant 2\sigma^2 \int |\nabla g|^2 d\mu \tag{8.26}$$

In this case, we write PI(σ^2) for short. Here, as usual,

$$\mathrm{Var}_\mu\left(g\right) = \int g^2 d\mu - \left(\int g d\mu\right)^2$$

stands for the variance of g and

$$\mathrm{Ent}_\mu\left(g^2\right) = \int g \log g d\mu - \int g d\mu \log \int g d\mu$$

denotes the entropy of $g \geq 0$ under the measure μ. We write LSI(σ^2). It is well known that LSI(σ^2) implies PI(σ^2).

These hypotheses are crucial in the study of the concentration of the spectral empirical distributions, especially of the linear functionals $\int f dF_n$ with individual smooth f on the line; See. for example [199, 444–446]. Qiu and Wicks [40] surveyed many recent results. A remarkable feature of the above approach to spectral analysis is that no specific knowledge about the non-explicit mapping from a random matrix to its spectral empirical measure is required. Instead, one may use general Lipschitz properties only, which are satisfied by this mapping. As for the general (not necessarily matrix) scheme, we shall only require hypotheses (8.25) and (8.26).

Theorem 8.4.1 Under PI(σ^2) on $\mathbb{R}^n$ ($n \geq 2$),

$$\mathbb{E}\int_{-\infty}^{\infty} \left|F_n(x) - F(x)\right| dx \leqslant C\sigma\left(\frac{M + \log n}{n}\right)^{1/3} \tag{8.27}$$

where $M = \frac{1}{\sigma}\max_{i,j}\left|\mathbb{E}X_i - \mathbb{E}X_j\right|$ and C is an absolute constant.

Note that the Poincare-type inequality (8.25) is invariant under shifts of the measure μ while the left-hand side of (8.27) is not. This is why the bound on the right-hand side of (8.27) should also depend on the means of the observations.

In terms of the ordered statistics $\tilde{X}_1 \leqslant \cdots \leqslant \tilde{X}_n$ of an arbitrary random vector $\mathbf{X} = \left(X_1, \ldots, X_n\right)$ in $\mathbb{R}^n$ there is a general two-sided estimate for the mean of the Kantorovich–Rubinstein distance:

$$\frac{1}{2n}\sum_{i=1}^{n}\left|\tilde{X}_i - \mathbb{E}\tilde{X}_i\right| \leqslant \mathbb{E}W_1\left(F_n, F\right) \leqslant \frac{2}{n}\sum_{i=1}^{n}\left|\tilde{X}_i - \mathbb{E}\tilde{X}_i\right| \tag{8.28}$$

Hence, under the conditions of Theorem 8.4.1, one may control the local fluctuations of $\tilde{X}_i$ (on average), which typically deviate from their mean by not more than $C\sigma\left(\frac{M+\log n}{n}\right)^{1/3}$.

Under a stronger hypothesis, such as (8.26), one can obtain more information about the fluctuations of $F_n(x) - F(x)$ for individual points x and thus gain some control of the Kolmogorov distance. Like the bound (8.27), such fluctuations will, on average, be shown to be at most

$$\beta = \frac{\left(\|F\|_{Lip}\sigma\right)^{2/3}}{n^{1/3}}$$

in the sense that

$$\mathbb{E}\left|F_n(x) - F(x)\right| \leqslant C\beta$$

where $\|F\|_{Lip}$, is the Lipschitz seminorm of F.

Theorem 8.4.2 Assume that F has a density, bounded by a number $\|F\|_{Lip}$, the Lipschitz seminorm of F. Under LSI(σ^2), for any $t > 0$

$$\mathbb{P}\left(\left\|F_n - F\right\| \geqslant t\right) \leqslant \frac{4}{t}e^{-c(t/\beta)^3} \tag{8.29}$$

In particular

$$\mathbb{E}\left\|F_n - F\right\| \leqslant C\beta\log^{1/3}\left(1 + \frac{1}{\beta}\right) \tag{8.30}$$

where c and C are positive absolute constants.

Example 8.4.3 (random matrix with Bernoulli random variables) Let all $X_i = \xi$, where ξ is uniformly distributed in [−1,1]. Here all random variables are identically distributed with $\mathbb{E}X_i = 0$. The joint distribution μ represents a uniform distribution on the main diagonal of the cube $[-1, 1]^n$, so it satisfies (8.25) and (8.26) with $\sigma = c\sqrt{n}$, where c is absolute. In this case, F is a uniform distribution on [−1,1], so $\|F\|_{Lip} = 1/2$ and β is of order 1. Hence, both sides of (8.30) are of order 1. □

8.4.1 Poincare-Type Inequalities, Tensorization

Poincare-type inequalities are known to hold with finite σ for many natural families of probability measures μ on $\mathbb{R}^n$. However, the problem of effective bounding of the Poincare constant σ^2 is not simple.

Definition 8.4.4 A probability measure μ on $\mathbb{R}^n$ satisfies a Poincare-type inequality with constant σ^2, $\sigma > 0$, called PI(σ^2), if for any bounded smooth function f on $\mathbb{R}^n$ with gradient ∇f

$$\mathrm{Var}_\mu\left(f\right) \leqslant \sigma^2 \int |\nabla f|^2 d\mu$$

Here, $\mathrm{Var}_\mu\left(f\right)$ is the variance of f under the measure μ, and $\sigma > 0$ is a constant depending on μ only. Note that the inequality itself is required to hold in the class of all bounded smooth functions $f : \mathbb{R}^n \to \mathbb{R}$. However, the smoothness of f may be relaxed to the property of being "locally Lipschitz," which means that near every point x the function f has a finite Lipschitz seminorm

$$\|f\|_{Lip} = \sup_{0<|x-y|<r} \frac{|f(x) - f(y)|}{|x - y|} \tag{8.31}$$

In this case the generalized modulus of the gradient

$$|\nabla g(x)| = \limsup_{y \to x} \frac{|g(x) - g(y)|}{|x - y|}$$

represents a finite Borel measurable function (see [447] for discussion and a general theory).

Now we state a very useful result that extends the Poincare-type inequality to product measures.

Lemma 8.4.5 Let $(\Omega, \mu) = \left(\Omega_1, \mu_1\right) \times \cdots \times \left(\Omega_n, \mu_n\right)$ and $f : \Omega \to \mathbb{R}$ be measurable. Denote f_i as function on Ω_i as defined by $f_i(x_i) = f\left(x_i, \ldots, x_n\right)$ where $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ are fixed. Then

$$\mathrm{Var}_\mu\left(f\right) \leqslant \sum_{i=1}^{n} \int \mathrm{Var}_{\mu_i}\left(f\right) d\mu$$

As an application of the lemma, we get the following well known theorem (see [446]):

Theorem 8.4.6 Let μ_i satisfy the Poincare-type inequality. Then

$$\mathrm{Var}_{\mu_i}\left(f\right) \leqslant \sigma^2 \int_{-\infty}^{\infty} \left|\nabla f_i(x)\right|^2 d\mu_i$$

for every bounded smooth function f_i on $\mathbb{R}$ with some constant σ^2. Then the product measure $\mu = \mu_1 \otimes \cdots \otimes \mu_1$ on the product space $\mathbb{R}^n$ satisfies the Poincare-type inequality. So for every bounded smooth function f on Ω

$$\mathrm{Var}_{\mu}\left(f\right) \leqslant \sigma^2 \int_{-\infty}^{\infty} |\nabla f(x)|^2 d\mu$$

Hence, the product measure also satisfies the Poincare-type inequality with the same constant σ^2.

The Poincare-type inequality is also stable under Lipschitz transformation.

Theorem 8.4.7 (Poincare-type inequality is stable under Lipschitz transformation) Let μ be a probability measure on $\mathbb{R}^n$ that satisfies a Poincare-type inequality with constant σ^2, called PI(σ^2). Suppose $T : \mathbb{R}^n \to \mathbb{R}^n$ is a Lipschitz transformation so that

$$\|T\mathbf{x} - T\mathbf{y}\|_{\mathbb{R}^k} \leqslant C\|\mathbf{x} - \mathbf{y}\|_{\mathbb{R}^n}$$

Let $\nu = T\mu^{-1}$. Then ν also satisfies a Poincare-type inequality with constant $(C\sigma)^2$. So we have

$$\mathrm{Var}_{\nu}\left(f\right) \leqslant C^2\sigma^2 \int |\nabla f|^2 d\nu$$

8.4.2 Empirical Poincare-Type Inequalities

Now, we are ready to consider Poincare-type inequalities for empirical measures to study the more general situation of an arbitrary random vector $\mathbf{X} = \left(X_1, \ldots, X_n\right)$ in $\mathbb{R}^n$ where $X_1, \ldots, X_n$ are not necessary independent or identically distributed. We consider F_n, the empirical distribution associated with the observations $X_1, \ldots, X_n$ and F the mean of the empirical distribution.

We want to measure the closeness of F_n to F. In general F is not continuous so it is hardly possible to work with the Kolmogorov distance $\rho(F_n, F)$ without additional assumptions on F (such as the existence and boundedness of its density). So, it seems more natural to choose weaker metrics, such as the Levy distance $L(F_n, F)$ or the Levy–Prokhorov distance $\pi(F_n, F)$, both of which are responsible for the weak convergence. Under moment assumptions, one may also involve the Kantorovich–Rubinstein distance

$$W_1\left(F_n, F\right) = \int_{-\infty}^{+\infty} \left|F_n(x) - F(x)\right| dx$$

Definition of Empirical Poincare-Type Inequality

By analytic hypotheses we mean integro-differential inequalities, imposed on the joint distribution μ of $\mathbf{X}$. As the simplest example, one may consider Poincare-type inequalities

$$\mathrm{Var}_{\mu}\left(f\right) \leqslant \sigma^2 \int |\nabla f|^2 d\mu$$

where $\mathrm{Var}_{\mu}\left(f\right)$ is the variance of f under the measure μ $\sigma > 0$ is a constant depending on μ only, and the inequality itself is required to hold in the class of all bounded smooth functions $f : \mathbb{R}^n \to \mathbb{R}$.

Let us apply the Poincare-type inequality to smooth functions of the form

$$f(\mathbf{x}) = \frac{g(x_1) + \cdots + g(x_n)}{n} = \int g dF_n, \quad \mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$$

where

$$F_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_k}$$

Then

$$|\nabla f(\mathbf{x})|^2 = \frac{g'(x_1)^2 + \cdots + g'(x_n)^2}{n^2} = \frac{1}{n} \int (g')^2 dF_n$$

As $\int F_n d\mu = F$, the Poincare-type inequality will take the form

$$\mathbb{E}\left| \int g dF_n - \int g dF \right|^2 \leqslant \frac{\sigma^2}{n} \int |g'|^2 dF$$

This inequality may be called an "empirical Poincare-type inequality." Note it remains to hold for complex-valued functions g, as well (by separating the real and imaginary parts of g).

Concentration of Empirical Characteristic Functions

The empirical Poincare-type inequality implies, for example

$$\mathbb{E}\left| \int g dF_n - \int g dF \right| \leqslant \frac{\sigma}{\sqrt{n}} \left(\int |g'|^2 dF \right)^{1/2}$$

So, linear functionals of the empirical measures, $\int g dF_n$ deviate from their mean $\int g dF$ on average at rate $\frac{1}{\sqrt{n}}$ like in the i.i.d. case—but under the additional assumption that g is smooth, such that the integral $\int |g'|^2 dF$ is finite.

In particular, we cannot apply it to an indicator function $g = 1_{(-\infty, x]}$ to get

$$\mathbb{E}\left| F_n(x) - F(x) \right| \leqslant \frac{C}{\sqrt{n}}$$

which is known to be true in the i.i.d. case. Nevertheless, at the expense of the rate, and properly changing the distance, one can suitably approximate indicator functions $g = 1_{(-\infty, x]}$ by smooth ones. The resulting bound should be weaker. Our problem is to estimate $\mathbb{E}\rho\left(F_n, F\right)$ in terms of σ^2 and n, where ρ is a given metric, responsible for the weak convergence on the real line. For example, it is known in [448] that if additionally $\mathbb{E}X_i = \mathbb{E}X_j$, for all i, j then

$$\mathbb{E}W_1\left(F_n, F\right) \leqslant C\sigma \left(\frac{\log(n+1)}{n} \right)^{1/3}$$

which is stated in Theorem 8.4.1.

Concentration of Empirical Distributions in Levy Metric

It is well known that closeness of characteristic functions in some sense implies closeness of the distributions with respect to the metrics that generate the topology of the weak convergence.

As an example, we will mention a result of Zolotarev about the Levy distance. Let F and G be distribution functions on the line with the characteristic functions

$$f(t)=\int_{-\infty}^{+\infty}e^{itx}dF(x),\quad g(t)=\int_{-\infty}^{+\infty}e^{itx}dG(x),\quad t\in\mathbb{R},$$

respectively. Recall that the Levy distance $L(F,G)$ is defined as the minimal value $h\geq 0$, such that

$$F(x-h)-h\leqslant G(x)\leqslant F(x+h)+h,\quad \text{for all } t\in\mathbb{R}$$

Theorem 8.4.8 (Zolotarev [449]) For any $T>0$

$$L(F,G)\leqslant c_1\int_0^T\frac{|f(t)-g(t)|}{t}dt+c_2\frac{\log(1+T)}{T}$$

where $c_1,c_2>0$ are universal constants.

See Appendix A of [450] for more details of this theorem. Note that there is no restriction on F and G. Hence, if f is close to g on a long interval $[0,T]$, then $L(F,G)$ will be small. The proofs of theorems rely heavily on this theorem.

Now we can apply Zolotarev's theorem to the empirical distribution functions to get

$$L\left(F_n,F\right)\leqslant c_1\int_0^T\frac{\left|f_n(t)-f(t)\right|}{t}dt+c_2\frac{\log(1+T)}{T}$$

where f_n is the characteristic function of F_n and f is the characteristic function of F. Taking the expectation and using Fubini's theorem, we obtain

$$\mathbb{E}L\left(F_n,F\right)\leqslant c_1\int_0^T\frac{\mathbb{E}\left|f_n(t)-f(t)\right|}{t}dt+c_2\frac{\log(1+T)}{T}\tag{8.32}$$

Now recall that an empirical Poincare-type inequality has the form

$$\mathbb{E}\left|\int gdF_n-\int gdF\right|^2\leqslant\frac{\sigma^2}{n}\int\left|g'(x)\right|^2dF(x)$$

and taking $g(x)=e^{itx}$ with parameter t gives

$$\int gdF_n=\frac{1}{n}\sum_{k=1}^{n}e^{itX_k}=f_n(t)\quad\text{and}\quad\int gdF=\int e^{itX}dF=f(t)$$

So with the above characteristic functions the empirical Poincare-type inequality implies we get

$$\mathbb{E}\left|f_n(t)-f(t)\right|\leqslant\sqrt{\mathbb{E}\left|f_n(t)-f(t)\right|^2}\leqslant\frac{\sigma|t|}{\sqrt{n}}$$

Substituting this back into (8.32), we can obtain the following inequality

$$\begin{aligned}\mathbb{E}L\left(F_n,F\right)&\leqslant c_1\int_0^T\frac{\sigma}{\sqrt{n}}dt+c_2\frac{\log(1+T)}{T}\\&=c_1\frac{\sigma T}{\sqrt{n}}+c_2\frac{\log(1+T)}{T},\quad\text{for } T>0\end{aligned}$$

In Zolotarev's bound, we may take constants $c_1 = 0.4$, and $c_2 = 4$. So we can consider following cases on the inequality

$$\mathbb{E}L\left(F_n, F\right) \leqslant C\left(\frac{\sigma T}{\sqrt{n}} + \frac{\log\left(1+T\right)}{T}\right)$$

Case I: $0 < \sigma \leq 1$. Then choosing $T = n^{1/4}$ gives

$$\mathbb{E}L\left(F_n, F\right) \leqslant C\frac{1+\log\left(1+n^{1/4}\right)}{n^{1/4}}$$

As $1+\log\left(1+n^{1/4}\right) \leqslant 5\log\left(n+1\right)$, we can obtain

$$\mathbb{E}L\left(F_n, F\right) \leqslant C\frac{\log\left(n+1\right)}{n^{1/4}}$$

for some universal constant C.

Case II: $\sigma > 1$. Then we choose $T = \frac{n^{1/4}}{\sqrt{\sigma}}$

$$\begin{aligned}\mathbb{E}L\left(F_n, F\right) &\leqslant \frac{n^{1/4}}{\sqrt{\sigma}}\left(1+\log\left(1+\frac{n^{1/4}}{\sqrt{\sigma}}\right)\right)\\ &\leqslant \frac{\sqrt{\sigma}}{n^{1/4}}\left(1+\log\left(1+n^{1/4}\right)\right)\\ &\leqslant C\frac{\sqrt{\sigma}}{n^{1/4}}\log\left(n+1\right),\end{aligned}$$

where C is some universal constant.

One can summarize these results as the following theorem:

Theorem 8.4.9 Let $\mathbf{X} = \left(X_1, \ldots, X_n\right)$ be a random vector in $\mathbb{R}^n$ and F_n be the empirical distribution associated with $\mathbf{X}$. Let $F = \mathbb{E}F_n$ and suppose $\mu = \mathcal{L}\left(\mathbf{X}\right)$ satisfies a Poincare-type inequality with constant σ^2. Then

1) if $0 \leq \sigma \leq 1$, then

$$\mathbb{E}L\left(F_n, F\right) \leqslant C\frac{\log\left(n+1\right)}{n^{1/4}}$$

2) if $\sigma > 1$, then

$$\mathbb{E}L\left(F_n, F\right) \leqslant C\frac{\sqrt{\sigma}\log\left(n+1\right)}{n^{1/4}}$$

The two cases in the theorem may be united by one inequality such as

$$\mathbb{E}L\left(F_n, F\right) \leqslant C\frac{\sqrt{1+\sigma}}{n^{1/4}}\log\left(n+1\right) \tag{8.33}$$

which holds for any σ (with some other constant C).

By similar methods, we can find a bound for higher moments. The L_p norm is defined as

$$\left\|L\left(F_n, F\right)\right\|_p = \left(\mathbb{E}\left(L\left(F_n, F\right)\right)^p\right)^{1/p}$$

For $p \geq 2$, after using the Zolotarev bound, we use Fubini's theorem. With some other constant $c > 0$, we obtain

$$c\left\|L\left(F_n, F\right)\right\|_p \leqslant \frac{p}{\sqrt{n}} T + \frac{\log(1+T)}{T} \tag{8.34}$$

Theorem 8.4.10 Let $\mathbf{X} = \left(X_1, \ldots, X_n\right)$ be a random vector in $\mathbb{R}^n$ and F_n be the empirical distribution associated with $\mathbf{X}$. Let $F = \mathbb{E}F_n$ and suppose $\mu = \mathcal{L}(\mathbf{X})$ satisfies a Poincare-type inequality with constant σ^2. Then for $p \geq 1$ we have

$$\left\|L\left(F_n, F\right)\right\|_p \leqslant C\sqrt{(1+\sigma)p}\frac{\log(n+1)}{n^{1/4}}$$

where C is an absolute constant.

When $p = 1$, we return to Theorem 8.4.9.

8.4.3 Concentration of Random Matrices

Now we study the empirical spectral measures F_n of the n ordered eigenvalues $X_1 \leqslant \cdots \leqslant X_n$ of a random symmetric matrix $\mathbf{M} = \left(\frac{1}{\sqrt{n}}\xi_{ij}\right), 1 \leqslant i, j \leqslant n$

$$\mathbf{M} = \frac{1}{\sqrt{n}}\begin{bmatrix} \xi_{11} & \xi_{12} & \cdots & \xi_{1n} \\ \xi_{21} & \xi_{22} & \cdots & \xi_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \xi_{n1} & \xi_{n2} & \cdots & \xi_{nn} \end{bmatrix}$$

for $\xi_{ij} = \xi_{ji}$ with *independent* entries above and on the diagonal ($n \geq 2$). Assume that $\mathbb{E}\xi_{ij} = 0$ and $\mathrm{Var}\left(\xi_{ij}\right) = 1$ so that the means $F = \mathbb{E}F_n$ converge to the semicircle law G with mean zero and variance one. The symmetry condition ensures that $\mathbf{M}$ has *real* eigenvalues eigenvalues $X_1 \leqslant \cdots \leqslant X_n$. The boundedness of moments of ξ_{ij} of any order will be guaranteed by (8.25). Consider $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}(x)$, the spectral empirical distribution associated with the particular values

$$X_1 = x_1, X_2 = X_2, \ldots, X_n = x_n$$

Theorem 8.4.11 If the distributions of the ξ_{ij}s satisfy the Poincare-type inequality $\mathrm{PI}(\sigma^2)$ on the real line, then

$$\mathbb{E}\int_{-\infty}^{\infty} \left|F_n(x) - F(x)\right| dx \leqslant C\sigma\frac{1}{n^{2/3}} \tag{8.35}$$

where C is an absolute constant. Moreover, under logarithmic Sobolev inequality $\mathrm{LSI}(\sigma^2)$

$$\mathbb{E}\left\|F_n(x) - G)\right\| \leqslant C\left(\frac{\sigma}{n}\right)^{2/3}\log^{1/3} n + \|F - G\| \tag{8.36}$$

By the convexity of the distance, we always have $\mathbb{E}\left\|F_n(x) - G)\right\| \geqslant \|F - G\|$. In some random matrix models, the Kolmogorov distance $\|F - G\|$ is known to tend to zero at

maximum rate of $1/n^{2/3-\epsilon}$. In the case of Gaussian ξ_{ij}, the distance $\|F-G\|$ is known to be of order $1/n$.

The distribution of the eigenvalues when ξ_{ij} are i.i.d. can be formulated as follows.

Theorem 8.4.12 (Wigner's semicircle law [451]) Let $\mathbf{M} = \left(\frac{1}{\sqrt{n}}\xi_{ij}\right), i \leq j$ be an $n \times n$ random symmetric matrix with eigenvalues $X_1 = x_1 \leqslant \cdots \leqslant X_n = x_n$. If (ξ_{ij}) are i.i.d., $\mathbb{E}\xi_{ij} = 0, \mathbb{E}\xi_{ij}^2 = 1$, and $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}(x)$, then

$$F = \mathbb{E}F_n(x) \Rightarrow G \quad \text{weakly}$$

where G is a distribution function of the semi-circle law with density

$$g(x) = \begin{cases} \frac{1}{2\pi}\sqrt{4-x^2}, & |x| \leqslant 2 \\ 0, & |x| > 2 \end{cases}$$

Now, we consider the case when ξ_{ij} may be *dependent* and not necessarily have same distributions. An important point to note in this case is that the joint distribution μ of the eigenvalues, as a probability measure on $\mathbb{R}^n$ represents the image of the joint distribution of ξ_{ij}s under a Lipschitz map T. We will use the following classical fact from the theory of matrix inequalities: Let $\mathbf{A} = \left(a_{ij}\right), \mathbf{B} = \left(b_{ij}\right), i \leq j$ be an $n \times n$ symmetric matrices with eigenvalues $x_1 \leqslant \cdots \leqslant x_n$ and $y_1 \leqslant \cdots \leqslant y_n$, respectively. Then

$$\sum_{i=1}^{n} \left(x_i - y_i\right)^2 \leqslant \sum_{i=1}^{n}\sum_{j=1}^{n} \left(a_{ij} - b_{ij}\right)^2$$

Thus, if $\mathbf{M} = \frac{1}{\sqrt{n}}\left(\xi_{ij}\right)_{i \leqslant j}$, in terms of our map

$$T : \mathbf{M} \mapsto \mathbf{x} = \left(x_1, \ldots, x_n\right) \in \mathbb{R}^n$$

we have

$$\begin{aligned} \left\|\mathbf{x} - \mathbf{x}'\right\|_{\mathbb{R}^n}^2 &= \left\|T\left(\mathbf{M}\right) - T\left(\mathbf{M}'\right)\right\|_{\mathbb{R}^n}^2 \\ &\leqslant \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\xi_{ij} - \xi_{ij}^{'}\right)^2 \\ &= \left\|T\left(\mathbf{M}\right) - T\left(\mathbf{M}'\right)\right\|_{HS}^2 \\ &\leqslant \frac{2}{n}\sum_{i \leqslant j}\left(\xi_{ij} - \xi_{ij}^{'}\right)^2 \end{aligned}$$

Therefore, we have the Lipschitz seminorm (defined in (8.31) below)

$$\|T\|_{Lip} \leqslant \sqrt{\frac{2}{n}}$$

Now, assume $\mathbf{M}$ is random and symmetric as before. Then $\mathbf{M} = \frac{1}{\sqrt{n}}\left(\xi_{ij}\right)_{i \leqslant j}$ can be viewed as a random vector in $\mathbb{R}^{n(n+1)/2}$ with distribution $Q = Q_\xi$. Assume that Q_ξ satisfies Poincare-type inequality PI $\left(\sigma^2\right)$ on $\mathbb{R}^{n(n+1)/2}$. Then, T pushes forward Q_ξ to $\mu = \mu_X$

on $\mathbb{R}^n$ and by Theorem 8.4.7, $\mu = Q_\xi T^{-1}$ satisfies PI $\left(\sigma_n^2\right)$ on $\mathbb{R}^n$ with $\sigma_n^2 = \frac{2}{n}\sigma^2$. Thus, we obtain

Theorem 8.4.13 Let $\mathbf{M} = \frac{1}{\sqrt{n}}\left(\xi_{ij}\right)_{i\leqslant j}$ be an $n\times n$ random symmetric matrix with eigenvalues $X_1 \leqslant \cdots \leqslant X_n$. Suppose that the joint distribution of ξ_{ij} satisfies Poincare-type inequality PI $\left(\sigma^2\right)$, then the empirical spectral distributions $F_n(x) = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}(x)$ satisfy the empirical Poincare-type inequality PI $\left(\sigma_n^2\right)$ $\mathbb{R}^n$ with $\sigma_n^2 = \frac{2}{n}\sigma^2$. So

$$\int g dF_n - \int g dF \leqslant \frac{2\sigma^2}{n}\int \left(g'\right)^2 dF$$

where, as before, $F = \mathbb{E}F_n(x)$.

Here, we study the concentration of F_n around F in terms of Levy distance. The L^p bound of the Levy distance in this situation using (8.34) will be

$$c\left\|L\left(F_n, F\right)\right\|_p \leqslant \frac{\sigma p}{n}t + \frac{\log(1+t)}{t}, \quad \text{for } t>0$$

where we have replaced σ with $\sigma\sqrt{2/n}$. We consider two cases again and choose t.

Case I: $0 \leq \sigma \leq 1$. Then choosing $t = \sqrt{n/p}$ gives

$$\begin{aligned} c\left\|L\left(F_n, F\right)\right\|_p &\leqslant \frac{\sigma p}{n}t + \frac{\log(1+t)}{t} \\ &\leqslant \sqrt{\frac{p}{n}}\left(1+\log\left(1+\sqrt{n}\right)\right) \\ &\leqslant C\sqrt{\frac{p}{n}}\log(n+1) \end{aligned}$$

Case II: $\sigma > 1$. Then choosing $t = \sqrt{n/\sigma p}$ gives

$$\begin{aligned} c\left\|L\left(F_n, F\right)\right\|_p &\leqslant \frac{\sigma p}{n}t + \frac{\log(1+t)}{t} \\ &\leqslant \sqrt{\frac{\sigma p}{n}}\left(1+\log\left(1+\sqrt{n}\right)\right) \\ &\leqslant C\sqrt{\sigma}\sqrt{\frac{p}{n}}\log(n+1) \end{aligned}$$

Therefore, the two cases can be combined in the following theorem.

Theorem 8.4.14 Let $\mathbf{M} = \frac{1}{\sqrt{n}}\left(\xi_{ij}\right)_{i\leqslant j}$ be an $n\times n$ random symmetric matrix with eigenvalues $X_1 \leqslant \cdots \leqslant X_n$. Suppose that the joint distribution of ξ_{ij} satisfies Poincare-type inequality PI $\left(\sigma^2\right)$, then the empirical spectral distributions

$$\left\|L\left(F_n, F\right)\right\|_p \leqslant C\sqrt{(1+\sigma)p}\frac{\log(n+1)}{\sqrt{n}}$$

for some absolute constant C where, $F_n(x)$ is the empirical spectral distribution, $F(x) = \mathbb{E}F_n(x)$, and L is the Levy distance.

Under the same assumption of Poincare-type inequality PI $\left(\sigma^2\right)$, we now have $n^{1/2}$ in the denominator rather than $n^{1/4}$ as in Theorem 8.4.10. However, $\sqrt{p}$ is still the same, so derivation of the deviation inequalities is not different from that of Theorem 8.4.10. In [450], the author followed the same procedure as before and apply Proposition B.3 of the Appendix section to Theorem 8.4.14. He obtained

$$\mathbb{E}\left\|L\left(F_n, F\right)\right\|_{\psi_2} \leqslant C\sqrt{(1+\sigma)}\frac{\log(n+1)}{\sqrt{n}}$$

in terms of the Orlicz norm generated by the Young function $\psi_2 = e^{t^2} - 1$. Let ψ be a Young function. For any measurable function Z on $\mathbb{R}$

$$\|Z\|_{\psi} = \inf\left\{\lambda > 0 : \mathbb{E}\psi\left(\frac{|Z|}{\lambda}\right) \leqslant 1\right\}$$

The Young function $\psi(t) = |t|^p$ in the examples yields the usual norm

$$\|Z\|_{\psi} = \left(\mathbb{E}|Z|^p\right)^{1/p} = \|Z\|_p \text{ on } L_p$$

Hence, by the definition of the Orlicz norm, we get

$$\mathbb{E}e^{L\left(F_n,F\right)^2/\alpha^2} \leqslant 2$$

where

$$\alpha = C\sqrt{(1+\sigma)}\frac{\log(n+1)}{\sqrt{n}}$$

Hence, by Chebyshev's inequality, we obtain the following deviation inequality.

Corollary 8.4.15 Under Poincare-type inequality PI $\left(\sigma^2\right)$, for any $t > 0$, we have

$$\mathbb{P}\left(L\left(F_n, F\right) > t\right) \leqslant 2e^{-t^2/\alpha^2}$$

where

$$\alpha = C\sqrt{(1+\sigma)}\frac{\log(n+1)}{\sqrt{n}}$$

and C is an absolute constant.

Therefore, in terms of the deviation inequality, we find the same Gaussian-type concentration. However, if we fix $t > 0$ and insert α we now obtain a fast decay as $n \to +\infty$ as

$$\mathbb{P}\left(L\left(F_n, F\right) > t\right) \leqslant 2e^{-Ct^2 n/\log^2(n+1)}$$

Under the assumption that the entries $\xi_{ij}, i \leq j$ are i.i.d., the concentration property of spectral empirical distributions were studied by many authors. In particular, assuming that each ξ_{ij} satisfies a log-Sobolev inequality with common constant σ^2, an analog of Corollary 8.4.15 can be found in a paper by Guionnet and Zeitouni [199].

Example 8.4.16 (new test metric for hypothesis testing in large random matrices) Consider the hypothesis test problem

$$\mathcal{H}_0 : \mathbf{A} = \frac{1}{\sqrt{n}}\left(\xi_{ij}\right)_{i\leqslant j}, 1 \leqslant i,j \leqslant n$$
$$\mathcal{H}_1 : \mathbf{B} \neq \frac{1}{\sqrt{n}}\left(\xi_{ij}\right)_{i\leqslant j}, 1 \leqslant i,j \leqslant n$$

where ξ_{ij} are independent may be dependent and not necessarily have same distributions such that Theorem 8.4.14 can be valid. Motivated by Theorem 8.4.14, we propose the use of the Levy distance as a new test metric:

$$\mathcal{H}_0 : \left\| L\left(F_n, F\right)\right\|_p \leqslant C\sqrt{(1+\sigma)p}\frac{\log(n+1)}{\sqrt{n}}$$
$$\mathcal{H}_1 : \left\| L\left(F_n, F\right)\right\|_p > C\sqrt{(1+\sigma)p}\frac{\log(n+1)}{\sqrt{n}}$$

where C is an absolute constant. Or even better, we use the Orlicz norm

$$\mathcal{H}_0 : \left\| L\left(F_n, F\right)\right\|_{\psi_2} \leqslant C\sqrt{(1+\sigma)}\frac{\log(n+1)}{\sqrt{n}}$$
$$\mathcal{H}_1 : \left\| L\left(F_n, F\right)\right\|_{\psi_2} > C\sqrt{(1+\sigma)}\frac{\log(n+1)}{\sqrt{n}}$$

□

The use of the Levy distance for test metric in a hypothesis test appears new, first here suggested by the author.

8.5 Random Quadratic Forms

Consider a quadratic form

$$\mathbf{y} = \mathbf{x}^H\mathbf{A}\mathbf{x}$$

where $\mathbf{x} = \left(X_1, \ldots, X_n\right)$ is, as usual, a random vector and $\mathbf{A} = \left(a_{ij}\right)_{1\leqslant i,j\leqslant n}$ a deterministic matrix. We say that X is sub-exponential with exponent α if there are constants $a, b > 0$ such that for all $t > 0$

$$\mathbb{P}\left(|X - \mathbb{E}X| \geqslant t^{\alpha}\right) \leqslant a\exp(-bt) \tag{8.37}$$

If $\alpha = 1/2$, then X is sub-Gaussian.

In 1971, Hanson and Wright [452] obtained the first important inequality for sub-Gaussian random variables. If $\mathbf{x} = \left(X_1, \ldots, X_n\right) \in \mathbb{R}^n$ be a random vector with X_i being i.i.d symmetric and sub-Gaussian random variables with mean 0 and variance 1. There exist constants $C, C' > 0$ (which may depend on the constants in (8.37)) such that the following holds. Let $\mathbf{A}$ be a real matrix of size n with entries a_{ij} and $\mathbf{B} := \left(\left|a_{ij}\right|\right)$. Then

$$\mathbb{P}\left(\left|\mathbf{x}^H\mathbf{A}\mathbf{x} - \operatorname{Tr}\mathbf{A}\right| \geqslant t\right) \leqslant C\exp\left(-C'\min\left\{\frac{t^2}{\|\mathbf{A}\|_F^2}, \frac{t}{\|\mathbf{B}\|_2^2}\right\}\right) \tag{8.38}$$

for any $t > 0$. Here $\|\mathbf{B}\|_F$ and $\|\mathbf{B}\|_2$ denote the Frobenius norm and the spectrum norm, respectively. Later Hsu, Kakade and Zhang [453] showed that one can obtain a better upper tail (notice that $\|\mathbf{B}\|_2^2$ is replaced by $\|\mathbf{A}\|_2^2$)

$$\mathbb{P}\left(\left|\mathbf{x}^H\mathbf{A}\mathbf{x} - \operatorname{Tr}\mathbf{A}\right| \geqslant t\right) \leqslant C\exp\left(-C'\min\left\{\frac{t^2}{\|\mathbf{A}\|_F^2}, \frac{t}{\|\mathbf{A}\|_2^2}\right\}\right) \tag{8.39}$$

under a considerably weaker assumption (which, in particular, does not require the X_i to be independent). For more recent results refer to [454].

8.6 Log-Determinant of Random Matrices

We say a random variable ξ satisfies condition **C0** (with positive constants C_1, C_2) if

$$\mathbb{P}\left(|\xi| \geqslant t\right) \leqslant C_1\exp\left(-t^{C_2}\right) \tag{8.40}$$

for all $t > 0$. Let $\mathbf{A}$ be an $n \times n$ random matrix whose entries are independent real random variables satisfying some natural conditions.

Theorem 8.6.1 ([455]) Assume that all atom variables a_{ij} of random matrix $\mathbf{A}$ of $n \times n$ satisfy condition **C0** with some positive constants C_1, C_2. Then

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\log|\det(\mathbf{A})| - \frac{1}{2}\log(n-1)!}{\sqrt{\frac{1}{2}\log n}} \leqslant x\right) - \Phi(x)\right| \leqslant \log^{-1/3+o(1)} n \tag{8.41}$$

Here, $\Phi(x) = \mathbb{P}\left(\mathcal{N}(0,1) < x\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(-t^2/2\right)dt$. The following form

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\log\det\left(\mathbf{A}^2\right) - \frac{1}{2}\log(n-1)!}{\sqrt{2\log n}} \leqslant x\right) - \Phi(x)\right| \leqslant \log^{-1/3+o(1)} n \tag{8.42}$$

is equivalent to (8.42). Using

$$\log\det(\cdot) = \operatorname{Tr}\log(\cdot)$$

we have

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\operatorname{Tr}\log\left(\mathbf{A}^2\right) - \frac{1}{2}\log(n-1)!}{\sqrt{2\log n}} \leqslant x\right) - \Phi(x)\right| \leqslant \log^{-1/3+o(1)} n \tag{8.43}$$

Consider a hypothesis testing problem

$$\begin{aligned}\mathcal{H}_0 &: \quad \mathbf{X}\\ \mathcal{H}_1 &: \quad \mathbf{X} + \mathbf{P}\end{aligned}$$

where $\mathbf{X}$ is a random matrix that is identical to $\mathbf{A}$ defined above, and $\mathbf{P}$ is the perturbation matrix. The above theorem can be used for this problem.

8.7 General MANOVA Matrices

Most material for Section 8.7 may be found in [456].

The three classical families of eigenvalue distributions of Gaussian random matrices are the Hermite, Laguerre and Jacobi ensembles. Hermite ensembles correspond to Wigner matrices, $\mathbf{X} = \mathbf{X}^H$; Laguerre ensembles describe sample covariance matrices, $\mathbf{X}\mathbf{X}^H$. The random $n \times n$ matrices yielding the Jacobi ensembles have the form of (8.45) for the special case when $\mathbf{X}$ and $\mathbf{Y}$ are Gaussian.

Consider the following matrix hypothesis problem

$$\begin{aligned} \mathcal{H}_0 &: \quad \mathbf{Y}\mathbf{Y}^H \\ \mathcal{H}_1 &: \quad \left(\mathbf{X}\mathbf{X}^H + \mathbf{Y}\mathbf{Y}^H\right)^{-1/2} \mathbf{Y}\mathbf{Y}^H \left(\mathbf{X}\mathbf{X}^H + \mathbf{Y}\mathbf{Y}^H\right)^{-1/2} \end{aligned} \tag{8.44}$$

If $\mathbf{X} = 0$, then the above two hypotheses are identical. As a result, the nonzero perturbation matrix $\mathbf{X}$ will make a difference in testing the two hypotheses.

The three classical families of eigenvalue distributions of Gaussian random matrices are the Hermite, Laguerre and Jacobi ensembles. Hermite ensembles correspond to Wigner matrices, $\mathbf{X} = \mathbf{X}^H$; Laguerre ensembles describe sample covariance matrices $\mathbf{X}\mathbf{X}^H$. The random $n \times n$ matrices yielding the Jacobi ensembles have the form

$$\left(\mathbf{X}\mathbf{X}^H + \mathbf{Y}\mathbf{Y}^H\right)^{-1/2} \mathbf{Y}\mathbf{Y}^H \left(\mathbf{X}\mathbf{X}^H + \mathbf{Y}\mathbf{Y}^H\right)^{-1/2} \tag{8.45}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are matrices of sizes $n \times [bn]$ and $n \times [an]$, respectively, with independent standard Gaussian entries. Here $a, b > 1$ are fixed parameters of the model, n is a large number, eventually tending to infinity, and $[\cdot]$ denotes the integer part. The matrix entries can be real, complex or self-dual quaternions, corresponding to the three symmetry classes, commonly distinguished by the parameter $\beta = 1, 2, 4$, respectively. The results presented in this section are insensitive to the symmetry class and for simplicity we will consider the complex case ($\beta = 2$).

Matrices of the form (8.45) are used in statistics for multivariate analysis of variance to determine correlation coefficients (Section 3.3 of [37]). This analysis is called MANOVA, though it has been largely limited to the special case when the entries of (8.45) are Gaussian.

In this section we address the case when the entries of X and Y in (8.45) are *independent* but have a general distribution with zero mean and unit variance. In particular, the matrix entries are not required to be identically distributed. We will call such a matrix with general entries a general MANOVA matrix.

Similarly to the Wigner and sample covariance matrices, the joint eigenvalue density of (8.45) is explicitly known only for the Gaussian case. When the entries are standard complex Gaussians, it is given by

$$\text{density}\left(\lambda_1, \ldots, \lambda_n\right) = C_{a,b,n} \prod_{i=1}^{n} \lambda_i^{(a-1)n} \left(1 - \lambda_i^{(b-1)n}\right) \prod_{1 \leqslant i < j \leqslant n} \left|\lambda_i - \lambda_j\right|^2 \tag{8.46}$$

where $C_{a,b,n}$ is a normalizing constant. The density has a similar form with different exponents when the matrix entries are real, or self-dual quaternions; see Section 3.6 of [62]. (8.46) defines the Jacobi ensemble, where the name refers to the form of the polynomial term in front of the Vandermonde determinant in (8.46).

The empirical density of the eigenvalues of (8.45)—equivalently, the one-point correlation function of (8.46)—converges almost certainly, as $n \to \infty$, to the distribution with density given by

$$f_M(x) = (a+b)\frac{\sqrt{(x-\lambda_-)(\lambda_+ - x)}}{2\pi x(1-x)} \cdot I_{[\lambda_-,\lambda_+]}(x) \tag{8.47}$$

where

$$\lambda_\pm = \left(\sqrt{\frac{a}{a+b}\left(1-\frac{1}{a+b}\right)} \pm \sqrt{\frac{1}{a+b}\left(1-\frac{a}{a+b}\right)}\right)^2 \tag{8.48}$$

The density f_M was determined by Wachter [457] and is discussed in Section 3.6 of [62]. Note that $\lambda_\pm \in (0,1)$, so that f_M is supported on a compact subinterval of (0, 1). We will refer to $f_M(x)$ as the *limiting distribution* of the eigenvalues of (8.45) or as the MANOVA distribution.

While the joint eigenvalue density (8.48) is valid only for the Gaussian case, the limiting empirical density is expected to be correct for general distributions as well, similarly to the universality of the Wigner semicircle law for Wigner matrices or the Marchenko–Pastur (MP) law for sample covariance matrices. Thus, general MANOVA matrices, the Jacobi ensemble and the distribution f_M constitute a triplet analogous to general Wigner matrices, the Hermite ensemble and the semicircle law or sample covariance matrices, the Laguerre ensemble and the Marchenko–Pastur law.

Given two positive constants $\gamma = (\gamma_1, \gamma_2)$, we say that a complex random variable Z is *γ-subexponential* if it satisfies the following conditions:

$$\begin{cases} \mathbb{E}Z = 0 \\ \mathbb{E}|Z|^2 = 1 \\ \mathbb{P}(|Z| \geqslant t^{\gamma_1}) \leqslant \gamma_2 e^{-t} \quad \text{for all } t > 0. \end{cases} \tag{8.49}$$

A set of random variables is uniformly γ-subexponential if each random variable is γ-subexponential for a common γ.

The main tool for this approach is the Stieltjes transform. Let $\mathbb{C}^+ = \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$. The Stieltjes transform of a real random variable with distribution function $F(x)$ is a function $\mathbb{C}^+ \to \mathbb{C}^+$ defined by

$$m(z) = \int \frac{1}{t-z} dF(t) \tag{8.50}$$

If the random variable has a density, then we also refer to the Stieltjes transform of the density. The Stieltjes transform of f_M is

$$m(z) = \frac{1}{2z(1-z)}\left[(2-a-b)z + a - 1 + \sqrt{(a+b)^2 z^2 - (a+b)\left(2(a+1) - \frac{a}{a+b}\right)z + (a-1)^2}\right] \tag{8.51}$$

For Hermitian matrices, we misuse notation and refer to the function

$$m(z) = \frac{1}{n}\mathrm{Tr}\,(\mathbf{A} - z\mathbf{I})^{-1}$$

as the Stieltjes transform of the Hermitian, $n \times n$ matrix $\mathbf{A}$. If $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $\mathbf{A}$, then we equivalently have

$$m_{\mathbf{A}}(z) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\lambda_i - z} = \frac{1}{n}\operatorname{Tr}(\mathbf{A} - z\mathbf{I})^{-1}$$

which is the Stieltjes transform of the empirical measure.

It is found that that the eigenvalues of the general MANOVA matrix behave close to what is indicated by $m_M(z)$ and f_M in the bulk with high probability.

Let λ_+ and λ_- be given in (8.48). Define

$$\mathcal{E}_{\kappa,\eta}^{(\lambda)} := \left\{E + i\eta \in \mathbb{C}^+ : E \in \left(\lambda_-, \lambda_+\right)\left(\lambda_+ - E\right)\left(E - \lambda_-\right) \geqslant \kappa\right\}$$

and set $\mathcal{E}_{\kappa}^{(\lambda)} = \mathcal{E}_{\kappa,0}^{(\lambda)}$.

Theorem 8.7.1 Fix two real parameters $a, b > 1$. Let $\mathbf{X}$ be an $n \times an$ random matrix and let $\mathbf{Y}$ be an $n \times bn$ random matrix independent of $\mathbf{X}$. We assume that both matrices have independent entries satisfying (8.49) for a common $\gamma = \left(\gamma_1, \gamma_2\right)$. Let $m_{n,M}(z)$ be the Stieltjes transform of the general MANOVA matrix

$$\left(\mathbf{X}\mathbf{X}^H + \mathbf{Y}\mathbf{Y}^H\right)^{-1/2}\mathbf{Y}\mathbf{Y}^H\left(\mathbf{X}\mathbf{X}^H + \mathbf{Y}\mathbf{Y}^H\right)^{-1/2} \tag{8.52}$$

- Then for any $\kappa, \eta > 0$ with $\eta > \frac{1}{n\kappa^2}\left(\log n\right)^{2C\log\log n}$, we have

$$\mathbb{P}\left(\sup_{z\in\mathcal{E}_{\kappa,\eta}^{(\lambda)}}\left|m_{n,\mathbf{M}}(z) - m_{\mathbf{M}}(z)\right| > \frac{\left(\log n\right)^{C\log\log n}}{\sqrt{\eta\kappa n}}\right) < n^{-c\log\log n} \tag{8.53}$$

  for all $n \geq n_0$ large enough and for constants $C, c > 0$. Here n_0, C and c depend only on γ.
- Let $N_\eta(E)$ denote the number of eigenvalues of (8.52) contained in $[E - \frac{\eta}{2}, E + \frac{\eta}{2}]$ and assume $\eta > \frac{1}{n\kappa^2}\left(\log n\right)^{3C\log\log n}$. Then

$$\mathbb{P}\left(\sup_{E\in\mathcal{E}_{\kappa}^{(\lambda)}}\left|\frac{N_\eta(E)}{n\eta} - f_{\mathbf{M}}(E)\right| > \frac{\left(\log n\right)^{C\log\log n}}{(\eta\kappa n)^{1/4}}\right) < n^{-c\log\log n}. \tag{8.54}$$

  We note that the entries of the matrices $\mathbf{X}$ and $\mathbf{Y}$ are not necessarily identically distributed.

For the hypothesis-testing problem (8.44), the idea is to transform the problem into another domain, using the Stieltjes transform

$$\begin{aligned}\mathcal{H}_0 &: m_{\mathbf{Y}\mathbf{Y}^H}(z)\\ \mathcal{H}_1 &: m_{\mathbf{M}}(z)\end{aligned} \tag{8.55}$$

As $n \to \infty$, both (8.44) and (8.55) tend to their individual nonrandom limits. If this is the case, hypothesis testing of two nonrandom functions can be easily handled. For example, we can study the test functions of $m_{\mathbf{Y}\mathbf{Y}^H}(z)$ and $m_M(z)$.

8.8 Finite Rank Perturbations of Large Random Matrices

In many applications, the $n \times m$ signal-plus-noise data or measurement matrix formed by stacking the m samples or measurements of $n \times 1$ observation vectors alongside each other can be modeled as

$$\mathbf{Y} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^H + \mathbf{X} \tag{8.56}$$

where $\mathbf{u}_i$ and $\mathbf{v}_i$ are left and right "signal" column vectors, σ_i are the associated "signal" values and $\mathbf{X}$ is the noise-only matrix of random noises. This model is ubiquitous in signal processing, statistics, and machine learning and is known under various guises as a signal subspace model [458], a latent variable statistical model [459], or a probabilistic PCA model [460].

The results presented in this section are very general in terms of possible distributions for the noise model $\mathbf{X}$, in a sense that will be made more precise shortly. Consider a particular case when $\mathbf{X}$ is Gaussian. The results in this section brings to light a general *principle*, which can be applied beyond the Gaussian case. Roughly speaking, this principle says that for $\mathbf{X}$ an $n \times m$ matrix (with $n, m >> 1$), if one adds an independent small rank perturbation $\sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^H$ to $\mathbf{X}$, then the extreme singular values will move to positions which are approximately the solutions z of the equations

$$\frac{1}{n} \operatorname{Tr} \frac{z}{z^2 \mathbf{I} - \mathbf{X}\mathbf{X}^H} \times \frac{1}{m} \operatorname{Tr} \frac{z}{z^2 \mathbf{I} - \mathbf{X}^H \mathbf{X}} = \frac{1}{\theta_i^2}, \quad (1 \leqslant i \leqslant r)$$

where we use the notation $\operatorname{Tr} \frac{1}{\mathbf{A}} = \operatorname{Tr} \mathbf{A}^{-1}$. In the case where these equations have no solutions (which means that the θ_i are below a certain threshold), then the extreme singular values of $\mathbf{X}$ will not move significantly. Similarly, we obtain the associated left and right singular vectors and give limit theorems for the fluctuations.

Let $\mathbf{X}_n$ be an $n \times m$ real or complex random matrix. Throughout this section, we assume that $n \leq m$ so that we may simplify the exposition of the proofs. We may do so without loss of generality because in a setting where $n > m$, the expressions derived will hold for $\mathbf{X}_n^H$. Recall that for $n \leq m$, the singular values of an $n \times m$ complex matrix $\mathbf{A}$ are the eigenvalues of the $n \times n$ matrix $\sqrt{\mathbf{A}\mathbf{A}^H}$. Let the $n \leq m$ singular values of $\mathbf{X}_n$ be sorted in nondecreasing order $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_n$. Let $\mu_{\mathbf{X}_n}(\cdot)$ be the empirical singular value distribution, with the probability measure defined as

$$\mu_{\mathbf{X}_n}(x) = \frac{1}{n} \sum_{i=1}^{n} \delta_{\sigma_i}(x)$$

Let m depend on n. We denote this dependence explicitly as m_n. Assume that as $n \to \infty$, $n/m_n \to c \in [0, 1]$.

Assumption 8.8.1 The probability measure $\mu_{\mathbf{X}_n}(\cdot)$ almost certainly coverges weakly to a nonrandom compactly supported probability measure $\mu_{\mathbf{X}}(\cdot)$.

When $\mathbf{X}_n$ has full rank (with high probability), the smallest singular value will be greater than zero.

Assumption 8.8.2 Let a be infimum of the support of $\mu_{\mathbf{X}_n}(\cdot)$. The smallest singular value of $\mathbf{X}_n$ converges almost certainly to a.

Assumption 8.8.3 Let b be supremum of the support of $\mu_{\mathbf{X}_n}(\cdot)$. The largest singular value of $\mathbf{X}_n$ converges almost certainly to b.

We shall consider the extreme singular values and the associated singular vectors of $\mathbf{Y}_n$, which is the random $n \times m$ matrix:

$$\mathbf{Y}_n = \mathbf{X}_n + \mathbf{A}_n \tag{8.57}$$

where the perturbation matrix $\mathbf{A}_n$ is defined below.

For a given $r \geq 1$, let $\theta_1 \geqslant \cdots \geqslant \theta_r > 0$ be deterministic nonzero real numbers, chosen independently of n. For every n, let $\mathbf{G}_u$, $\mathbf{G}_v$ be two independent matrices with sizes respectively $n \times r$ and $m \times r$ with i.i.d. entries distributed according to a fixed probability measure ν on $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$. We introduce the column vectors $\mathbf{u}_1, \ldots, \mathbf{u}_r \in \mathbb{K}^{n\times 1}$ and $\mathbf{v}_1, \ldots, \mathbf{v}_r \in \mathbb{K}^{m\times 1}$ obtained from $\mathbf{G}_u$, $\mathbf{G}_v$ by either:

- The *i.i.d. model.* Setting $\mathbf{u}_i$ and $\mathbf{v}_i$ to equal the i-th column of $\frac{1}{\sqrt{n}}$ and $\mathbf{G}_u \frac{1}{\sqrt{m}} \mathbf{G}_v$, respectively or,
- The *Orthonormalized model.* Setting $\mathbf{u}_i$ and $\mathbf{v}_i$ to equal vectors obtained from a Gram-Schmidt (or QR factorization) of $\mathbf{G}_u$ and $\mathbf{G}_v$, respectively.

We define the random perturbing matrix $\mathbf{A}_n \in \mathbb{K}^{n\times m}$ as

$$\mathbf{A}_n = \sum_{i=1}^{r} \theta_i \mathbf{u}_i \mathbf{v}_i^H$$

In the orthonormalized model, the θ_is are the nonzero singular values of $\mathbf{A}_n$ and the $\mathbf{u}_i$s and $\mathbf{v}_i$s are the left and right associated singular vectors.

Assumption 8.8.4 The probability measure ν has mean zero, variance one and that satisfies a log-Sobolev inequality.

See [49] for the treatment of log-Sobolev inequality. We make several remarks. First, if ν is the standard real or complex Gaussian distribution, then the singular vectors produced using the orthonormalized model will have uniform distribution on the set of r orthogonal random vectors. Second, if $\mathbf{X}_n$ is random but has a bi-unitarily *invariant* distribution and $\mathbf{A}_n$ is nonrandom with rank r, then we are in the same setting as the orthonormalized model for the results that follow. More generally, our idea in defining both of our models (the i.i.d. one and the orthonormalized one) was to show that if $\mathbf{A}_n$ is chosen independently from $\mathbf{X}_n$ in a somehow "isotropic way," i.e. via a distribution that is not far away from being invariant by the action of the orthogonal group by conjugation, then a BBP phase transition [335] occurs, which is governed by a certain integral transform of the limit empirical singular value distribution of $\mathbf{X}_n$, namely $\mu_{\mathbf{X}_n}(x)$. Third, the framework could easily be adapted to the case where the distribution of the entries of $\mathbf{G}_u$ and the distribution of the entries of $\mathbf{G}_v$ are not the same, both satisfying Assumption 8.8.4.

In Theorems 8.8.5, we suppose Assumptions 8.8.1, 8.8.3, and 8.8.4 are valid.

For a function f and $t \in \mathbb{R}$, we use the notation

$$f\left(t^{+}\right) = \lim_{z \downarrow t} f(z); \qquad f\left(t^{-}\right) = \lim_{z \uparrow t} f(z)$$

We define θ_c, the critical threshold of the phase transition, as

$$\theta_c := \frac{1}{\sqrt{D_{\mu_X}(b^{+})}}$$

with the convention that $(+\infty)^{-1/2} = 0$, and where $D_{\mu_X}(\cdot)$, the D-transform of the measure μ_X is the function, depending on c, defined by

$$D_{\mu_X}(z) = \left[\int \frac{z}{z^2 - t^2} d\mu_X\right] \times \left[c \int \frac{z}{z^2 - t^2} d\mu_X + \frac{1-c}{z}\right] \quad \text{for } z > b$$

In the theorems below, $D^{-1}_{\mu_X}(\cdot)$ will denote its functional inverse on $[b, +\infty)$. We use notation $\xrightarrow{a.s.}$ denote almost sure convergence.

Theorem 8.8.5 Largest singular value phase transition. The r largest singular values of the $n \times m$ perturbed matrix $\mathbf{Y}_n = \mathbf{X}_n + \mathbf{A}_n$ exhibit the following behavior as $n, m_n \to \infty$ and $n/m_n \to c$. We have that for each fixed $1 \le i \le r$

$$\sigma_i\left(\mathbf{X}_n + \mathbf{A}_n\right) \xrightarrow{a.s.} \begin{cases} D^{-1}_{\mu_X}\left(1/\theta_i\right) & \text{if } \theta_i > \theta_c \\ b & \text{otherwise} \end{cases} \tag{8.58}$$

Moreover, for each fixed $i > r$, we have that $\sigma_i\left(\mathbf{X}_n + \mathbf{A}_n\right) \xrightarrow{a.s.} b$.

The D-transform in free probability theory is critical. The C-transform with ratio c of a probability measure μ on $\mathbb{R}^{+}$, defined as

$$C_\mu(z) = U\left(z\left(D^{-1}_\mu(z)\right)^2 - 1\right)$$

where

$$U(z) = \begin{cases} \frac{-c-1+\left[(c+1)^2+4cz\right]^{1/2}}{2c} & \text{when } c > 0 \\ z & \text{when } c = 0 \end{cases}$$

is the analog of the logarithm of the Fourier transform for the rectangular free convolution with ratio c (see [461, 462] for an introduction to the theory of rectangular free convolution) in the sense described next.

Let $\mathbf{A}_n$ and $\mathbf{B}_n$ be independent $n \times m$ rectangular random matrices that are invariant, in law, by conjugation by any orthogonal (or unitary) matrix. Suppose that, as $n, m \to \infty$ with $n/m \to c$, the empirical singular value distributions $\mu_{\mathbf{A}_n}$ and $\mu_{\mathbf{B}_n}$ of $\mathbf{A}_n$ and $\mathbf{B}_n$ satisfy $\mu_{\mathbf{A}_n} \to \mu_{\mathbf{A}}$, and $\mu_{\mathbf{B}_n} \to \mu_{\mathbf{B}}$. Then by [463] the empirical singular value distribution $\mu_{\mathbf{A}_n + \mathbf{B}_n}$ of $\mathbf{A}_n + \mathbf{B}_n$ satisfy

$$\mu_{\mathbf{A}_n + \mathbf{B}_n} \to \mu_{\mathbf{A}} \boxplus \mu_{\mathbf{B}}$$

where $\mu_{\mathbf{A}} \boxplus \mu_{\mathbf{B}}$ is a probability measure which can be characterized in terms of the C-transform as

$$C_{\mu_{\mathbf{A}} \boxplus \mu_{\mathbf{B}}}(z) = C_{\mu_{\mathbf{A}}}(z) + C_{\mu_{\mathbf{B}}}(z)$$

The coefficients of the series expansion of $U(z)$ are the rectangular free cumulants with ratio c of μ (see [463] for an introduction to the rectangular free cumulants). The connection between free rectangular additive convolution and $D_{\mu}^{-1}(z)$ (via the C-transform) and the appearance of $D_{\mu}^{-1}(z)$ in Theorem 8.8.5 could be of independent interest to free probabilists: the emergence of this transform in the study of isolated singular values completes the picture of [464], where the transforms linearizing additive and multiplicative free convolutions.

Example 8.8.6 (Gaussian rectangular random matrices with nonzero mean) Let $\mathbf{X}_n$ be an $n \times m$ real (or complex) matrix with independent, zero mean, normally distributed entries with variance $1/m$. It is known [163,172] that, as $n, m \to \infty$ with $n/m \to c \in (0,1]$, the spectral measure of the singular values of $\mathbf{X}_n$ converges to the distribution with density

$$d\mu_{\mathbf{X}}(x) = \frac{1}{\pi c}\frac{1}{x}\sqrt{4c - (x^2 - 1 - c)^2}\mathbb{I}_{(a,b)}(x)\,dx$$

where $a = 1 - \sqrt{c}$ and $b = 1 + \sqrt{c}$ are the end points of the support of $\mu_{\mathbf{X}}$. It is known [163] that the extreme eigenvalues converge to the bounds of this support.

Associated with this singular measure, we have, after some manipulation

$$D_{\mu_{\mathbf{X}}}^{-1}(z) = \sqrt{\frac{(z+1)(cz+1)}{z}},$$
$$D_{\mu_{\mathbf{X}}}(z) = \frac{z^2 - (c+1) - \sqrt{(z^2 - (c+1))^2 - 4c}}{2c}, \quad D_{\mu_{\mathbf{X}}}\left(b^+\right) = \frac{1}{\sqrt{c}}$$

Thus for any $n \times m$ deterministic matrix $\mathbf{A}_n$ with r nonzero singular values $\theta_1 \geqslant \cdots \geqslant \theta_r > 0$ (r independent of n, m), for any fixed $i \geq 1$, by Theorem 8.8.5, we have

$$\sigma_i\left(\mathbf{X}_n + \mathbf{A}_n\right) \xrightarrow{a.s.} \begin{cases} \sqrt{\frac{(1+\theta_i^2)(c+\theta_i^2)}{\theta_i^2}} & \text{if i} \leqslant r \text{ and } \theta_i > c^{1/4} \\ 1 + \sqrt{c} & \text{otherwise} \end{cases}$$

as $n \to \infty$. For the i.i.d. model defined above, this formula allows us to recover some of the results of [335]. Now, let us turn our attention to the singular vectors. Let $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{v}}$ be left and right unit singular vectors of $\mathbf{X}_n + \mathbf{A}_n$. In the setting where $r = 1$, let $\mathbf{A}_n = \theta \mathbf{u}\mathbf{v}^H$. Then using Theorems 2.10 and 2.11 in [367], we have

$$|\langle \tilde{\mathbf{u}}, \mathbf{u} \rangle|^2 \xrightarrow{a.s.} \begin{cases} 1 - \frac{c\left(1+\theta^2\right)}{\theta^2(\theta^2+c)} & \text{if } \theta \geqslant c^{1/4} \\ 0 & \text{otherwise} \end{cases}$$

where $\langle \tilde{\mathbf{u}}, \mathbf{u} \rangle$ is the inner product of the true leading eigenvector and the corresponding perturbed leading eigenvector. The phase transitions for the eigenvectors of $\left(\mathbf{X}_n + \mathbf{A}_n\right)^H \left(\mathbf{X}_n + \mathbf{A}_n\right)$ or for the pairs of singular vectors of $\mathbf{X}_n + \mathbf{A}_n$ can be similarly computed to yield the expression

$$|\langle \tilde{\mathbf{v}}, \mathbf{v} \rangle|^2 \xrightarrow{a.s.} \begin{cases} 1 - \frac{\left(c+\theta^2\right)}{\theta^2(\theta^2+1)} & \text{if } \theta \geqslant c^{1/4} \\ 0 & \text{otherwise} \end{cases}$$

□

Example 8.8.7 (square Haar unitary matrices) Let $\mathbf{X}_n$ be a Haar distributed unitary (or orthogonal) random matrix. All of its singular values are equal to one, so that it has limiting spectral measure

$$\mu_{\mathbf{X}}(x) = \delta_1,$$

with $a = b = 1$ being the end points of the support of $\mu_{\mathbf{X}}$. Associated with this spectral measure, we obtain (of course, $c = 1$)

$$D_{\mu_{\mathbf{X}}}(z) = \frac{z^2}{(z^2-1)^2}, \quad \text{for } z \geqslant 0, z \neq 1$$

thus for $\theta > 0$

$$D^{-1}_{\mu_{\mathbf{X}}}\left(1/\theta^2\right) = \begin{cases} \frac{\theta+\sqrt{\theta^2+4}}{2} & \text{if the inverse is computed on } (1,+\infty) \\ \frac{-\theta+\sqrt{\theta^2+4}}{2} & \text{if the inverse is computed on } (0,1) \end{cases}$$

Thus for any $n \times n$, rank r perturbing matrix $\mathbf{A}_n$ with r nonzero singular values $\theta_1 \geqslant \cdots \geqslant \theta_r > 0$ where neither r, nor the θ_is depend on n, for any fixed $i = 1, \ldots, r$, by Theorem 8.8.5 we have

$$\sigma_i\left(\mathbf{X}_n + \mathbf{A}_n\right) \xrightarrow{a.s.} \frac{\theta_i + \sqrt{\theta_i^2+4}}{2} \text{ and } \sigma_{n+1-i}\left(\mathbf{X}_n + \mathbf{A}_n\right) \xrightarrow{a.s.} \frac{-\theta_i + \sqrt{\theta_i^2+4}}{2}$$

while for any fixed $i > r+1$, both $\sigma_i\left(\mathbf{X}_n + \mathbf{A}_n\right)$ and $\sigma_{n+1-i}\left(\mathbf{X}_n + \mathbf{A}_n\right) \xrightarrow{a.s.} 1$. □

8.8.1 Non-asymptoic, Finite-Sample Theory

The spiked true covariance matrix was first considered by the seminar paper [177], which is highly cited. The spiked true covariance matrix $\mathbf{\Sigma}$ has all but a few eigenvalues equal one:

$$\mathbf{\Sigma} \sim \text{diag}\left\{{\theta_1}^2, \ldots, \theta_{r+s}^2, 1, \ldots, 1\right\} \in \mathbb{R}^{p \times p}$$

where

$$\theta_1 \geqslant \cdots \geqslant \theta_r > 1 > \theta_{r+1} \geqslant \cdots \geqslant \theta_{r+s} > 0$$

where there are $(p-r)$ eigenvalues that are less than one and there are s nonzero eigenvalues. Let $\mathbf{X} \in \mathbb{R}^{p \times n}$ with entries X_{ij} being i.i.d. $\mathcal{N}(0,1)$. Consider the sample covariance matrix

$$\mathbf{S}_n = \frac{1}{n}\left(\mathbf{\Sigma}^{1/2}\mathbf{X}\right)\left(\mathbf{\Sigma}^{1/2}\mathbf{X}\right)^T$$

and

$$\lambda_1\left(\mathbf{S}_n\right) \geqslant \cdots \geqslant \lambda_p\left(\mathbf{S}_n\right)$$

are eigenvalues sorted in a nondecreasing order. If $\mathbf{\Sigma} = \mathbf{I}_p$, then $\mathbf{S}_n$ is a Wishart matrix. So the spiked true covariance matrix model can be considered as a finite rank perturbation of the Wishart matrix ensemble.

Our results about the spiked population model are divided into two parts. Theorem 3.2 of [465] established deviation bounds for the largest eigenvalues, and Theorem 3.3

of [465] established deviation bounds for the smallest eigenvalues. We can summarize these two theorems as the following.

Let θ^2 be an eigenvalue of the true covariance matrix $\mathbf{\Sigma}$, $\theta^2 \neq 1$. Then the corresponding "spiked eigenvalue" $\lambda\left(\mathbf{S}_n\right) \in \mathbb{R}^{p\times p}$ of the sample covariance matrix will satisfy

$$\mathbb{P}\left(\left|\lambda\left(\mathbf{S}_n\right) - \lambda_{\theta,c}\right| > t\right) \leqslant C_1 e^{-C_2 n t^2} \tag{8.59}$$

where $\lambda_{\theta,c}$ is defined as

$$\lambda_{\theta,c} = \begin{cases} \theta^2 + c \cdot \frac{\theta^2}{\theta^2 - 1} & \text{if } \theta^2 > 1 + \sqrt{c}, \text{or } c < 1, \theta^2 < 1 - \sqrt{c} \\ \left(1 + \sqrt{c}\right)^2 & \text{if } 1 < \theta^2 \leqslant 1 + \sqrt{c} \\ \left(1 - \sqrt{c}\right)^2 & \text{if } c < 1, 1 - \sqrt{c} \leqslant \theta^2 \leqslant 1 \end{cases}$$

where $c = \frac{p-r}{n}$. The right-hand side of (8.59) is Gaussian with variance proportion to $1/\sqrt{n}$. The proof of (8.59) is based on the concentration of measure phenomenon, which has a large literature, such as [40, 49, 446]. The approach of using the Stieltjes transform is to study the asymptotic limits, as $n \to \infty$. On the other hand, (8.59) applies to the nonasymptotic case when n is large but finite. In practice, we can apply (8.59) to moderate data size n.

8.9 Hypothesis Tests for High-Dimensional Datasets

This section analyzes whether standard covariance matrix tests work when dimensionality is large, and in particular larger than sample size. In the latter case, the singularity of the sample covariance matrix makes likelihood ratio tests (LRT) degenerate, but other tests based on quadratic forms of sample covariance matrix eigenvalues remain well defined. Previous authors have noted that the LRT may not perform well in finite samples.

Since the mid-1990s, the practical environment has changed dramatically, with the spectacular evolution of data-acquisition technologies and computing facilities. At the same time, applications have emerged in which the number of experimental units is comparatively small but the underlying dimension is massive [466]. Ideas from random matrix theory are connected with large covariance matrices. The most informative components for inference may or may not be the principal components [467].

Data visualization [468] is important. Visual statistical methods are used with an inferential framework and protocol, modeled on confirmatory statistical testing. In this framework, plots take on the role of test statistics, and human cognition the role of statistical tests. Statistical significance of "discoveries" is measured by having the human viewer compare the plot of the real data set with collections of plots of simulated data sets.

Many empirical problems involve large-dimensional covariance matrices. Sometimes the dimensionality p is even larger than the sample size n, which makes the sample covariance matrix $\mathbf{S}$ singular. For concreteness, we focus on two common testing problems: (i) the covariance matrix $\mathbf{\Sigma}$ is proportional to the identity $\mathbf{I}$ (sphericity):

$$\mathcal{H}_0 : \mathbf{\Sigma} = \sigma^2 \mathbf{I} \quad \text{vs.} \quad \mathcal{H}_1 : \mathbf{\Sigma} \neq \sigma^2 \mathbf{I}$$

where σ^2 is unspecified. (ii) the covariance matrix $\mathbf{\Sigma}$ is equal to the identity $\mathbf{I}$ (sphericity):

$$\mathcal{H}_0 : \mathbf{\Sigma} = \mathbf{I} \quad \text{vs.} \quad \mathcal{H}_1 : \mathbf{\Sigma} \neq \mathbf{I}$$

The identity $\mathbf{I}$ can be replaced with any other matrix $\mathbf{\Sigma}_0$ by multiplying the data by $\mathbf{\Sigma}_0^{-1/2}$. For both hypotheses the likelihood ratio test statistic is degenerate when p exceeds n : $p > n$. This steers us toward other test statistics that do not degenerate, such as

$$U = \frac{1}{p} \operatorname{Tr} \left[\frac{\mathbf{S}}{(1/p) \operatorname{Tr}(\mathbf{S})} - \mathbf{I} \right]^2 \quad \text{and} \quad V = \frac{1}{p} \operatorname{Tr} \left[(\mathbf{S} - \mathbf{I})^2 \right] \tag{8.60}$$

The asymptotic framework assumes n goes to infinity while p remains fixed. It treats terms of order p/n like terms of order $1/n$, which is inappropriate if p is of the same order of magnitude as n. The robustness of tests based on U and V against high dimensionality was studied for the first time by Ledoit and Wolf [469].

We study the asymptotic behavior of U and V as p and n go to infinity together with the ratio p/n converging to a limit $y \in (0, +\infty)$. The singular case corresponds to $y > 1$. The robustness issue boils down to power and size: is the test still consistent? Is the n-limiting distribution under the null still a good approximation? Surprisingly, we find opposite answers for U and V. The power and the size of the sphericity test based on U turn out to be robust against p large, and even larger than n. But the test of $\mathcal{H}_0 : \mathbf{\Sigma} = \mathbf{I}$ based on V is not consistent against every alternative when p goes to infinity with n, and its n-limiting distribution differs from its (n, p)-limiting distribution under the null. This prompts us to introduce the modified statistic

$$W = \frac{1}{p} \operatorname{Tr} \left[(\mathbf{S} - \mathbf{I})^2 \right] - \frac{p}{n} \left[\frac{1}{p} \operatorname{Tr}(\mathbf{S}) \right]^2 + \frac{p}{n} \tag{8.61}$$

Note that W only involves diagonal elements of the sample covariance matrix $\mathbf{S}$ through the trace function.

The maximal invariant likelihood ratio test asymptotically has good power in the spiked covariance matrix model, whereas the standard likelihood ratio test has no power at all [470].

8.9.1 Motivation for Likelihood Ratio Test (LRT) and Covariance Matrix Tests

Traditional statistical theory, particularly in multivariate analysis, did not contemplate the demands of high dimensionality [95, 177] in data analysis. The classical multivariate analysis textbooks [37, 371] were developed under the assumption that the dimension of the dataset, conventionally denoted by p, is considered a fixed small constant or is at least negligible compared with the sample size n. Because their dimensions can be proportionally large compared with the sample size, this assumption, however, is no longer true for many modern datasets, such as smart grid data, financial data, consumer data, manufacturing data, and multimedia data.

In classic statistical inference, the likelihood ratio test (LRT) is one widely used method for hypothesis testing. An *advantage* of using the LRT is that one does not have to estimate the variance of the test statistics. It is well known that the asymptotic distribution of the LRT is chi-square under certain regularity conditions when the dimension p is a small constant or is negligible compared with the sample size n. However, the chi-square approximation does not fit the distribution of the LRT

very well for the high-dimension case, especially when p grows with the sample size n.

The failure of the traditional multivariate method for high-dimensional data had been observed by Dempster [471] in as early as 1958. Bai and Saranadasa [472] did some further work. Bai *et al.* [160] studied the likelihood ratio test (LRT) for the covariance matrix of a normal distribution and showed that using the traditional chi-square approximation to the limiting distribution of the test statistic will result in a much inflated test size (or alpha error) even with moderate size of p and n. They developed corrections to the traditional likelihood ratio test to make it suitable for testing a high-dimensional normal distribution $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with

$$\mathcal{H}_0 : \boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_p \quad \text{vs.} \quad \mathcal{H}_1 : \boldsymbol{\Sigma} \neq \sigma^2 \mathbf{I}_p$$

The test statistic is chosen to be

$$L_n = \text{Tr}(\mathbf{S}) - \log\det(\mathbf{S}) - p$$

where $\mathbf{S}$ is the sample covariance matrix from the data. In their derivation, the dimension p is no longer considered a fixed constant, but rather a variable that goes to infinity along with the sample size n; and the ratio between p and n converges to a constant y:

$$\lim_{n\to\infty} \frac{p_n}{n} = y \in (0, 1) \tag{8.62}$$

Jiang *et al.* [473] further extend Bai's result to cover the case of $y = 1$. Jiang and Yang [474] studied several other classical likelihood ratio tests for means and covariance matrices of high-dimensional normal distributions. Most of these tests have the asymptotic results for their test statistics derived decades ago under the assumption of a large n but a fixed p. Their results supplement these traditional results in providing alternatives to analyze high-dimensional datasets including the critical case $p/n \to 1$.

The central limit theorems of the LRT statistics mentioned in Jiang and Yang [474] the context of $\lim_{n\to\infty} \frac{p}{n} = y \in (0, 1)$ are new in the literature. Similar results are Bai *et al.* [160] and Jiang *et al.* [473]. The methods of the proofs in the three papers are different: the random matrix theory is used in Bai *et al.* [160]; the Selberg integral is used in Jiang *et al.* [473]. Jiang and Yang [474] obtained the central limit theorems by analyzing the moments of the LRT statistics.

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be independent and identically distributed p-dimensional random vectors with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. We form a random matrix $\mathbf{X}$ of $n \times p$. Testing the covariance matrix

$$\mathcal{H}_0 : \boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_p \quad \text{vs.} \quad \mathcal{H}_1 : \boldsymbol{\Sigma} \neq \sigma^2 \mathbf{I}_p \tag{8.63}$$

where $\mathbf{I}_p$ is the p-dimensional identity matrix and σ^2 is an unknown but finite positive constant. The identity hypothesis in (8.102) covers the hypothesis for

$$\mathcal{H}_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 \quad \text{vs.} \quad \mathcal{H}_1 : \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}_0$$

for an *arbitrary* specifically known invertible covariance matrix $\boldsymbol{\Sigma}_0$. This comment is true to all the tests. For convenience, we often deals with (8.63).

The traditional method based on the sample covariance $\mathbf{X}'\mathbf{X}$ such as the likelihood ratio test, see Anderson [371], can no longer function when $p \to \infty$ as $n \to \infty$. Almost all statistical theories dealing with large samples were developed through probabilistic

limiting theorems of fixed dimension p and increasing sample size n. Modern random matrix theory, however, predicts that, when the dimension of $\mathbf{x}_i$, p is not negligible with respect to the sample size n, the sample covariance matrix $\mathbf{S}$ as a function of n

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^H = \frac{1}{n}\mathbf{X}\mathbf{X}^H$$

does not approach $\mathbf{\Sigma}$ of $p \times p$. Therefore, classical statistical procedures based on an approximation of $\mathbf{\Sigma}$ by $\mathbf{S}$ become *inconsistent* or very *inefficient* in situations with high-dimensional data. There is thus a great need to develop new statistical tools for high-dimensional data analysis [53]. Estimation of $\mathbf{\Sigma}$ is among the central problems in high-dimensional statistics, with applications including principal component analysis, Kalman filtering and independent component analysis.

When dimension p and the sample size n, are comparable, i.e. $n/p \to c \in (0, \infty)$, many methods were developed based on random matrix theory [35]. By assuming $\boldsymbol{\mu} = 0$, the largest eigenvalue has been considered for testing hypothesis in (8.63) by Johnstone [177] in the Gaussian case, and by Péché [475] in the more general case where the distribution is assumed to be sub-Gaussian tails. Ledoit and Wolf [469] first used the trace of the quadratic forms of the sample covariance as a new test statistic to test the null hypothesis under the normality assumption. By weakening the conditions, Chen, Zhang and Zhong [476] also introduced a similar test statistic.

Besides the likelihood ratio tests, many other traditional hypothesis tests in multivariate analysis have also been revisited in the past decade for high-dimensional cases. Examples include Srivastava [477,478] and Schott [479–482]. Fujikoshi *et al.* [483] gave a book-length survey on multivariate methods under the high-dimensional framework when $\lim_{n\to\infty} \frac{p}{n} = y > 0$. Cai and Ma [484] optimal hypothesis testing for high dimensional covariance matrices. Wang, Cao and Miao [485] deal with asymptotic power of likelihood ratio tests for high-dimensional data.

8.9.2 Estimation of Covariance Matrices Using Loss Functions

To take advantage of some prior information about $\mathbf{\Sigma}$, we can use the loss functions to estimate the true covariance matrix. We examine some properties of these loss functions.

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in $\mathbb{C}^p$, be independent identically distributed p-dimensional normal vectors with mean 0 and common *unknown* nonsingular covariance matrix $\mathbf{\Sigma}$. From the joint probability density function

$$\mathbb{P}_{\mathbf{\Sigma}} = \frac{1}{(2\pi)^{np/2}(\det\mathbf{\Sigma})^{n/2}}\exp\left(-\frac{1}{2}\operatorname{Tr}\mathbf{\Sigma}^{-1}\mathbf{x}^H\mathbf{x}\right) \tag{8.64}$$

of the random vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$, it is easy to see that $\mathbf{S}$

$$\mathbf{S} = \sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^H = \mathbf{X}\mathbf{X}^H$$

is a sufficient statistic, $\mathbf{X} = \left(\mathbf{x}_1, \ldots, \mathbf{x}_n\right)$, and that the maximum likelihood estimator of the covariance matrix $\mathbf{\Sigma}$ is

$$\hat{\mathbf{\Sigma}}^{ML}(\mathbf{S}) = \frac{1}{n}\mathbf{S}$$

We will describe some estimators $\hat{\mathbf{\Sigma}}(\mathbf{S})$, which are better than the maximum likelihood estimator with respect to the loss function

$$L\left(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}\right) = \operatorname{Tr} \mathbf{\Sigma}^{-1}\hat{\mathbf{\Sigma}} - \log\det \mathbf{\Sigma}^{-1}\hat{\mathbf{\Sigma}} - p \tag{8.65}$$

We will use the loss function defined by (8.65) largely because it is comparatively easy to work with this loss function. However, it also has all the appealing properties of loss functions

- $L\left(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}\right) \geqslant 0$, with equality if and only if $\mathbf{\Sigma} = \hat{\mathbf{\Sigma}}$.
- $L\left(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}\right)$ a convex function of its second argument;
- $L\left(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}\right)$ invariant under linear transformations of $\mathbb{R}^n$, i.e., for any nonsingular $p \times p$ matrix $\mathbf{A}$,

$$L\left(\mathbf{A}\mathbf{\Sigma}\mathbf{A}^H, \mathbf{A}\hat{\mathbf{\Sigma}}\mathbf{A}^H\right) = L\left(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}\right). \tag{8.66}$$

Let us recall the theorem on diagonalization of any Hermitian matrix by orthogonal transformations.

For any Hermitian $p \times p$ matrix $\mathbf{H}$ there exist a unique diagonal matrix $\mathbf{D}$ and an orthogonal matrix $\mathbf{V}$ such that

$$\mathbf{H} = \mathbf{V}\mathbf{D}\mathbf{V}^H$$

and the diagonal elements d_i of $\mathbf{D}$ satisfy the inequalities

$$d_1 \geqslant d_2 \geqslant \cdots \geqslant d_p \geqslant 0$$

If the matrix H is positive definite, then $d_p > 0$. By the same theorem, the covariance matrix $\mathbf{\Sigma}$ and the observed sample matrix $\mathbf{S}$ are representable in the form

$$\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H, \quad \mathbf{S} = \tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{U}}^H$$

where λ_i is the i-th element of the diagonal matrix $\mathbf{\Lambda}$ and $\tilde{\lambda}_i$ is the i-th element of the diagonal matrix $\tilde{\mathbf{\Lambda}}$. Both $\mathbf{U}$ and $\tilde{\mathbf{U}}$ are orthogonal matrices and

$$\lambda_1 \leqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p > 0, \quad \tilde{\lambda}_1 \leqslant \tilde{\lambda}_2 \geqslant \cdots \geqslant \tilde{\lambda}_p > 0$$

It can be shown that, if p is sufficiently large and $\mathbf{\Sigma}$ is close to the unit matrix, while i/p and $1 - i/p$ are sufficiently small, then the i-th eigenvalue $\tilde{\lambda}_i/n$ of the matrix $\mathbf{S}/n$ are not likely to be close to the λ_i. Furthermore, the ratio $\tilde{\lambda}_1/\tilde{\lambda}_p$ is likely to be much greater than the ratio λ_1/λ_p. This suggests that instead of the traditional estimator

$$\frac{1}{n}\mathbf{S} = \frac{1}{n}\tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{U}}^H$$

it is better to use an estimator of the form

$$\hat{\mathbf{\Sigma}} = \mathbf{U}\varphi\left(\tilde{\mathbf{\Lambda}}\right)\mathbf{U}^H, \tag{8.67}$$

where φ is an appropriately chosen function mapping the space of positive diagonal matrices onto itself. The function φ should be chosen so that the ratio $\varphi_1\left(\tilde{\mathbf{\Lambda}}\right)/\varphi_p\left(\tilde{\mathbf{\Lambda}}\right)$ of the first and the last diagonal elements of the matrix $\varphi\left(\tilde{\mathbf{\Lambda}}\right)$ is considerably smaller than the ratio $\tilde{\lambda}_1/\tilde{\lambda}_p$.

Let me describe briefly the theoretical results for the case when $\mathbf{\Sigma} = \mathbf{I}$, which has been most extensively studied. A similar problem was considered by the physicist Wigner before statisticians became interested in the problem. In particular, the following theorem was proved by Wigner.

Theorem 8.9.1 As $p \to \infty$ with $n/p \to y > 1$, the empirical distribution function of the eigenvalues $\tilde{\lambda}_1/n, \ldots, \tilde{\lambda}_p/n$ converges in probability to the nonrandom function

$$F(x) = c \int_a^x \frac{1}{t} \sqrt{(t-a)(b-t)} dt, \quad a \leqslant t \leqslant b$$

where

$$a = \left(1 - \frac{1}{y}\right)^2; \quad b = \left(1 + \frac{1}{y}\right)^2$$

When estimating the inverse $\mathbf{\Sigma}^{-1}$, it may be better to use $\left[\hat{\mathbf{\Sigma}}(\mathbf{S})\right]^{-1}$ instead of $\left[\mathbf{S}/n\right]^{-1}$, because

$$\mathbb{E}\left[(\mathbf{S}/n)^{-1}\right] = \frac{n}{n-p-1}\mathbf{\Sigma}^{-1}$$

whence it follows that if $n - p - 1$ is small, the diagonal elements of $(\mathbf{S}/n)^{-1}$ are always greater than the elements of $\mathbf{\Sigma}^{-1}$.

Recall that if $\mathbf{A}$ is an $n \times n$ Hermitian matrix then there exists $\mathbf{V}$ unitary and $\mathbf{D} = \operatorname{diag}\left(d_1, \ldots, d_n\right)$ such that $\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{V}^H$. Given a continuous function f we define $f(\mathbf{A})$ as

$$f(\mathbf{A}) = \mathbf{V} \operatorname{diag}\left(f\left(d_1\right), \ldots, f\left(d_n\right)\right) \mathbf{V}^H$$

To return to the estimators $\hat{\mathbf{\Sigma}} = \mathbf{U}\varphi\left(\tilde{\mathbf{\Lambda}}\right)\mathbf{U}^H$, let us consider the choice of the function ϕ. With

$$\psi_i\left(\tilde{\mathbf{\Lambda}}\right) = \frac{1}{\lambda_i}\varphi_i\left(\tilde{\mathbf{\Lambda}}\right), \quad i = 1, \ldots, p$$

we have the risk function of an estimator $\hat{\mathbf{\Sigma}} = \mathbf{U}\varphi\left(\tilde{\mathbf{\Lambda}}\right)\mathbf{U}^H$:

$$\begin{aligned}
\mathbb{E}_{\mathbf{\Sigma}}\left\{L\left(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}\right)\right\} &= \mathbb{E}_{\mathbf{\Sigma}}\left\{\operatorname{Tr}\mathbf{\Sigma}^{-1}\hat{\mathbf{\Sigma}} - \log\det\mathbf{\Sigma}^{-1}\hat{\mathbf{\Sigma}} - p\right\} \\
&= \mathbb{E}_{\mathbf{\Sigma}}\left\{(n-p+1)\sum_{k=1}^{n}\psi_k\left(\tilde{\mathbf{\Lambda}}\right) - \sum_{k=1}^{n}\log\psi_k\left(\tilde{\mathbf{\Lambda}}\right) + \right. \\
&\quad - 2\sum_{j=1}^{p}\sum_{i>j}^{p}\frac{\tilde{\lambda}_j\psi_j\left(\tilde{\mathbf{\Lambda}}\right) - \tilde{\lambda}_i\psi_i\left(\tilde{\mathbf{\Lambda}}\right)}{\tilde{\lambda}_j - \tilde{\lambda}_i} \\
&\quad \left. + 2\sum_{j=1}^{p}\tilde{\lambda}_j\frac{\partial}{\partial\tilde{\lambda}_j}\psi_j\left(\tilde{\mathbf{\Lambda}}\right) - \sum_{j=1}^{p}\log\chi^2_{n-j+1} - p,\right\}
\end{aligned} \tag{8.68}$$

See Stein [486] for the proof of (8.68).

If we choose ψ_i so as to minimize (8.68), ignoring the effect of $\sum_{j=1}^{p} \tilde{\lambda}_j \frac{\partial}{\partial \tilde{\lambda}_j} \psi_j\left(\tilde{\mathbf{\Lambda}}\right)$, we obtain

$$\varphi_j^{(i)}\left(\tilde{\mathbf{\Lambda}}\right) = \frac{\tilde{\lambda}_j}{\alpha_j\left(\tilde{\mathbf{\Lambda}}\right)}, \quad j = 1, \ldots, p$$

where

$$\alpha_i\left(\tilde{\mathbf{\Lambda}}\right) = n + p - 2j + 1 + 2\sum_{i>j} \frac{\tilde{\lambda}_i}{\tilde{\lambda}_j - \tilde{\lambda}_i} - 2\sum_{i<j} \frac{\tilde{\lambda}_i}{\tilde{\lambda}_j - \tilde{\lambda}_i}$$

These $\varphi_1^{(i)}\left(\tilde{\mathbf{\Lambda}}\right)$ often turn out to vary excessively with the index. In particular, it happens frequently that they do not satisfy $\varphi_1^{(i)}\left(\tilde{\mathbf{\Lambda}}\right) \geqslant \cdots \geqslant \varphi_p^{(i)}\left(\tilde{\mathbf{\Lambda}}\right)$ and sometimes some of the $\varphi_1^{(i)}\left(\tilde{\mathbf{\Lambda}}\right)$ are even negative. Fairly reasonable estimators are obtained by defining

$$\varphi_j^{(2)}\left(\tilde{\mathbf{\Lambda}}\right) = \frac{\sum_{i\in\Omega_i} \tilde{\lambda}_i}{\sum_{i\in\Omega_i} \alpha_i\left(\tilde{\mathbf{\Lambda}}\right)}$$

where Ω_i is the set of consecutive integers such that

$$j \in \Omega_j$$
$$i \in \Omega_j \Leftrightarrow j \in \Omega_i$$

and

$$\varphi_1^{(2)}\left(\tilde{\mathbf{\Lambda}}\right) \geqslant \varphi_2^{(2)}\left(\tilde{\mathbf{\Lambda}}\right) \geqslant \cdots \geqslant \varphi_p^{(2)}\left(\tilde{\mathbf{\Lambda}}\right)$$

For every j, the set Ω_j depends on $\tilde{\mathbf{\Lambda}}$ and among all the sets Ω having the properties of the set Ω_j the latter is the smallest.

Example 8.9.2 (estimator invariant under the linear transformation) Consider the problem in which we observe $\mathbf{x}_1, \ldots, \mathbf{x}_n$ independently normally distributed p-dimensional random vectors with mean 0 and unknown covariance matrix $\mathbf{\Sigma}$ where $n \geq p$. Suppose we want to estimate $\mathbf{\Sigma}$, say by $\hat{\mathbf{\Sigma}}$, with loss (distance) function

$$L\left(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}\right) = \operatorname{Tr} \mathbf{\Sigma}^{-1}\hat{\mathbf{\Sigma}} - \log\det \mathbf{\Sigma}^{-1}\hat{\mathbf{\Sigma}} - p \tag{8.69}$$

The problem is invariant under the transformations

$$\mathbf{x}_i \to \mathbf{A}\mathbf{x}_i, \quad \mathbf{\Sigma} \to \mathbf{A}\mathbf{\Sigma}\mathbf{A}^H, \quad \hat{\mathbf{\Sigma}} \to \mathbf{A}\hat{\mathbf{\Sigma}}\mathbf{A}^H$$

where $\mathbf{A}$ is an arbitrary nonsingular $p \times p$ matrix. Also

$$\mathbf{S} = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^H$$

is a sufficient statistic and if we make the transformation $\mathbf{x}_i \to \mathbf{A}\mathbf{x}_i$, then $\mathbf{S} \to \mathbf{A}\mathbf{S}\mathbf{A}^H$. We may confine our attention to estimators that are functions of $\mathbf{S}$ alone. The condition of invariance of an estimator φ (a function on the set of positive definite $p \times p$ Hermitian matrices itself) under transformation by the matrix $\mathbf{A}$ is

$$\varphi\left(\mathbf{A}\mathbf{T}\mathbf{A}^H\right) = \mathbf{A}\varphi\left(\mathbf{T}\right)\mathbf{A}^H \quad \text{for all } \mathbf{T} \tag{8.70}$$

We shall find that this $\varphi(\mathbf{S})$ is not a scalar multiple of $\mathbf{S}$. Similar results hold for the quadratic loss (distance) function

$$L_0\left(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}\right) = \mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}} - \mathbf{I}\right)^2$$

Putting $\mathbf{T} = I$ in (8.70) we find

$$\varphi\left(\mathbf{A}\mathbf{A}^H\right) = \mathbf{A}\varphi(\mathbf{I})\,\mathbf{A}^H.$$

When we let $\mathbf{A}$ range over the set of diagonal matrices with ± 1 on the diagonal, this gives

$$\varphi(\mathbf{I}) = \mathbf{A}\varphi(\mathbf{I})\,\mathbf{A}^H \tag{8.71}$$

which implies that $\varphi(\mathbf{I})$ is a diagonal matrix, say $\boldsymbol{\Delta}$, with i-th diagonal element Δ_i. This, together with (8.70), determines φ since any positive definite Hermitian matrix $\mathbf{S}$ can be factored as $\mathbf{S} = KK^H$ with $\mathbf{K}$ lower triangular (with positive diagonal elements) and we then have

$$\varphi(\mathbf{S}) = \mathbf{K}\varphi(\boldsymbol{\Delta})\,\mathbf{K}^H \tag{8.72}$$

Since the group of lower triangular matrices operates transitively on the parameter space, the risk of an invariant procedure φ is constant. Thus we compute the risk only for $\boldsymbol{\Sigma} = \mathbf{I}$. Then we have

$$\begin{aligned}\rho(\mathbf{I}, \varphi(\mathbf{S})) &= \mathbb{E}\left[\mathrm{Tr}\,\varphi(\mathbf{S}) - \log\det\varphi(\mathbf{S}) - p\right] \\ &= \mathbb{E}\left[\mathrm{Tr}\,\mathbf{K}\boldsymbol{\Delta}\mathbf{K}^H - \log\det\mathbf{K}\boldsymbol{\Delta}\mathbf{K}^H - p\right] \\ &= \mathbb{E}\,\mathrm{Tr}\,\mathbf{K}\boldsymbol{\Delta}\mathbf{K}^H - \log\det\boldsymbol{\Delta} - \mathbb{E}\log\det\mathbf{S} - p\end{aligned} \tag{8.73}$$

But

$$\begin{aligned}\mathbb{E}\,\mathrm{Tr}\,\mathbf{K}\boldsymbol{\Delta}\mathbf{K}^H &= \sum_{i,j}\Delta_i\mathbb{E}K_{ij}^2 \\ &= \sum\Delta_i\mathbb{E}\chi_{n-i+1+p-i}^2 = \sum(n+p-2i+1)\,\Delta_i\end{aligned} \tag{8.74}$$

since the elements of $\mathbf{K}$ are independent of each other, the i-th diagonal element being distributed as χ_{n-i+1} and the elements below the diagonal normal with mean 0 and variance 1. Also, for the same reason

$$\mathbb{E}\log\det\mathbf{S} = \sum_{i=1}^{p}\mathbb{E}\log\chi_{n-i+1}^2 \tag{8.75}$$

It follows that

$$\begin{aligned}\rho(\boldsymbol{\Sigma}, \varphi(\mathbf{S})) &= \rho(\mathbf{I}, \varphi(\mathbf{S})) \\ &= \sum_{i=1}^{p}\left[(n+p-2i+1)\,\Delta_i - \log\Delta_i\right] - \sum_{i=1}^{p}\mathbb{E}\log\chi_{n-i+1}^2 - p\end{aligned} \tag{8.76}$$

This attains its minimum value of

$$\begin{aligned}\rho(\boldsymbol{\Sigma}, \varphi^*(\mathbf{S})) &= \sum_{i=1}^{p}\left[1 - \log\frac{1}{n+p-2i+1} - \mathbb{E}\log\chi_{n-i+1}^2\right] - p \\ &= \sum\left[\log(n+p-2i+1) - \mathbb{E}\log\chi_{n-i+1}^2\right]\end{aligned} \tag{8.77}$$

when

$$\Delta_i = \frac{1}{n+p-2i+1} \tag{8.78}$$

We have thus found the minimax estimator in a class of estimators that includes the natural estimators (multiples of $\mathbf{S}$) to be different from the natural estimators. □

Example 8.9.3 (relationship between two loss functions) The two loss functions are defined as

$$L_1\left(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}\right) = \operatorname{Tr} \boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}} - \log\det \boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}} - p, \quad L_2\left(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}\right) = \operatorname{Tr}\left(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}} - \mathbf{I}\right) \tag{8.79}$$

We define the risk function as

$$R_i\left(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}\right) \equiv \mathbb{E}\left[L_i\left(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}\right) | \boldsymbol{\Sigma}\right], \qquad i = 1, 2$$

Let ε be a real number, and $\mathbf{A}$ a $p \times p$ Hermitian matrix. If we apply the expansion

$$\log\det(\mathbf{I} + \varepsilon\mathbf{A}) = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} \varepsilon^k \operatorname{Tr}\left(\mathbf{A}^k\right) \tag{8.80}$$

to L_1, then a relationship is obtained between L_1 and L_2. The series converges if the spectral radius of $\mathbf{A}$ is less than unity and the radius of convergence $0 \leq \varepsilon \leq 1$. In particular, set $\varepsilon = 1$, factor $\boldsymbol{\Sigma}^{-1}$ as $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Omega}^2$, and expand

$$\begin{aligned}
\log\det\left(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}\right) &= \log\det\left(\boldsymbol{\Omega}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Omega}\right) \\
&= \log\det(\mathbf{I} + \mathbf{A}) \quad \left(\mathbf{A} = \boldsymbol{\Omega}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Omega} - \mathbf{I}\right) \\
&= \operatorname{Tr}(\mathbf{A}) - \frac{1}{2}\operatorname{Tr}\left(\mathbf{A}^2\right) + \frac{1}{3}\operatorname{Tr}\left(\mathbf{A}^3\right) - \cdots + \frac{(-1)^{k-1}}{k}\operatorname{Tr}\left(\mathbf{A}^k\right) + \cdots \\
&= \operatorname{Tr}\left(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1} - \mathbf{I}\right) - \frac{1}{2}\operatorname{Tr}\left(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}} - \mathbf{I}\right)^2 + \cdots
\end{aligned} \tag{8.81}$$

From (8.81), loss functions can be written as

$$L_1\left(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}\right) = \frac{1}{2}L_2\left(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}\right) - \frac{1}{3}\operatorname{Tr}\left(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}} - \mathbf{I}\right)^3 + \cdots \tag{8.82}$$

so it is plausible that estimators which perform well (mod L_1) also perform well (mod L_2). □

8.9.3 Covariance Matrix Tests

As pointed out above, the likelihood ratio tests cannot be used when the sample size n is smaller than the dimension p. The singularity of the sample covariance matrix makes likelihood ratio tests degenerate but other tests based on sample covariance matrix remain well defined. We use the so-called moment method [67].

Roughly, we have three methods to deal with large-dimensional random matrices: the moment method (Section 3.12), the Stieltjes transform method (Section 3.13) and the logarithmic potential method[1].

1 The logarithmic potential method has been used first by Tao and Vu [326] to prove the circular law. See [328, 329] for details.

Recall from (3.53) that, for a positive integer k, the k-th moment of the empirical spectral density is given by

$$m_k = \int x^k F_{\mathbf{S}}(dx) = \frac{1}{N} \operatorname{Tr}\left(\mathbf{S}^k\right) = \frac{1}{n} \operatorname{Tr}\left(\left(\frac{1}{n}\mathbf{X}^H\mathbf{X}\right)^k\right) \tag{8.83}$$

This expression plays a fundamental role in random matrix theory. By the moment convergence theorem, the problem of showing that the expected ESD of a sequence of random matrices $\mathbf{S} = \frac{1}{n}\mathbf{X}^H\mathbf{X}$ tends to a limit reduces to showing that, for each fixed k, the sequence

$$\frac{1}{n}\mathbb{E} \operatorname{Tr}\left(\left(\frac{1}{n}\mathbf{X}^H\mathbf{X}\right)^k\right)$$

tends to a limit. We know that when the matrix sizes $n \times p$ of $\mathbf{X}$ goes large together, m_k will reach their limits. The proof of the convergence of the ESD $F_{\mathbf{X}^H\mathbf{X}/n}$ to a limit usually reduces to the estimation of the second or higher moments

$$\frac{1}{n} \operatorname{Tr}\left(\left(\frac{1}{n}\mathbf{X}^H\mathbf{X}\right)^k\right)$$

From (8.83), we have confidence in solely studying the statistics of the moments.

The basic idea behind a family of algorithms consists of finding the map

$$\theta \mapsto m_k(\theta) = \int x^k dF(x)$$

that links the parameter θ of the observation model, the moments of the limiting Marchenko–Pastur distribution. Because the sample moments $\hat{m}_k$

$$\hat{m}_k = \frac{1}{p} \operatorname{Tr} \mathbf{S}^k = \frac{1}{p} \sum_{i=1}^{p} \lambda_i^k(\mathbf{S}), \quad k = 1, \ldots, q$$

are *consistent estimators* of m_k, it is then natural to use the moment method for the inference of the parameters θ.

Section 8.9.3

Define the true moments as

$$Y_i = (1/p) \operatorname{Tr} \mathbf{\Sigma}^i, \quad i = 1, \ldots, 8$$

We make the following assumptions:

- (A) As $p \to \infty$, $Y_i \to Y_i^0$, $0 < Y_i^0 < \infty, i = 1, \ldots, 8$.
- (B) $n = O\left(p^\delta\right)$, $0 < \delta < 1$.

Under assumption (A), and as $n \to \infty$, an unbiased and consistent estimators of Y_1 and Y_2 are respectively given by

$$\hat{Y}_1 = \frac{1}{p} \operatorname{Tr}(\mathbf{S}) \tag{8.84}$$

and

$$\hat{Y}_2 = \frac{n^2}{(n-1)(n+2)} \frac{1}{p} \left[\operatorname{Tr}\left(\mathbf{S}^2\right) - \frac{1}{n}(\operatorname{Tr} \mathbf{S})^2\right] \tag{8.85}$$

From the definition of $\hat{Y}_1$ and $\hat{Y}_2$, it follows that

$$\frac{1}{p}\operatorname{Tr}\mathbf{S}^2 = \hat{Y}_2 + \frac{1}{pn}(\operatorname{Tr}\mathbf{S})^2 = \hat{Y}_2 + \frac{p}{n}\hat{Y}_1^2$$

Thus, unless p/n goes to zero as $n \to \infty$ and $p \to \infty$, $(\operatorname{Tr}\mathbf{S})^2/p$ is not a *consistent* estimator of $(1/p)\operatorname{Tr}\boldsymbol{\Sigma}^2$, while $\hat{Y}_2$ is always a consistent estimator of Y_2 irrespective of how $n \to \infty$, provided the assumption (A) is satisfied.

It can be shown that asymptotically,

$$\begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}, \frac{1}{np}\begin{pmatrix} 2Y_2 & 4Y_3 \\ 4Y_2 & 8Y_4 + 4(p/n)Y_2^2 \end{pmatrix}\right] \tag{8.86}$$

Let $n \to \infty$ and $p \to \infty$ such that $p/n \to c$. Then, asymptotically

$$\begin{pmatrix} \frac{1}{p}\operatorname{Tr}\mathbf{S} \\ \frac{1}{p}\operatorname{Tr}\mathbf{S}^2 \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} Y_1 \\ Y_2 + cY_1^2 \end{pmatrix}, \frac{1}{n^2 c}\Delta\right] \tag{8.87}$$

where

$$\Delta = \begin{pmatrix} 2Y_2 & 4\left(cY_1Y_2 + Y_3\right) \\ 4\left(cY_1Y_2 + Y_3\right) & 4\left(2Y_2 + cY_2^2 + 4cY_1Y_3 + 2c^2Y_1^2Y_2\right) \end{pmatrix}$$

Now we are in a position to study a test for the sphericity. We consider the problem of testing the hypothesis

$$\mathcal{H}_0 : \boldsymbol{\Sigma} = \sigma^2\mathbf{I} \quad \text{vs.} \quad \mathcal{H}_1 : \boldsymbol{\Sigma} \neq \sigma^2\mathbf{I} \tag{8.88}$$

when the sample size $n+1$, $\mathbf{x}_1, \ldots, \mathbf{x}_{n+1}$ is drawn from $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. When $n > p$, the most appropriate commonly used test statistic is the likelihood ratio test, which has been shown in [487] to have a monotone power function. However, when $n < p$ the likelihood ratio test is not available. Here we consider a test based on a consistent estimator of a parametric function of $\boldsymbol{\Sigma}$, which separates the null hypothesis from the alternative hypothesis.

As with the likelihood ratio test in classical multivariate statistics, the testing problem remains invariant under the transformation $\mathbf{x} \to \mathbf{G}\mathbf{x}$, where $\mathbf{G}$ belongs to the group of orthogonal matrices. The problem also remains invariant under the scalar transformation $\mathbf{x} \to c\mathbf{x}$. Thus, we may assume without any loss of generality that

$$\boldsymbol{\Sigma} = \operatorname{diag}\left(\lambda_1, \ldots, \lambda_p\right) \tag{8.89}$$

a $p \times p$ diagonal matrix. From the Cauchy–Schwarz inequality, it follows that

$$\left(\sum_{i=1}^{p} \lambda_i \times 1\right)^2 \leqslant p\sum_{i=1}^{p}\lambda_i^2$$

with equality holding if and only if $\lambda_i \equiv c$ for some constant c. Thus

$$\gamma \equiv \frac{\sum_{i=1}^{p}\lambda_i^2/p}{\left(\sum_{i=1}^{p}\lambda_i/p\right)^2} \geqslant 1 \tag{8.90}$$

and is equal to 1 if and only if $\lambda_i \equiv c$ for some constant c. Hence, we may consider testing the hypothesis

$$\mathcal{H}_0 : \gamma - 1 = 0 \quad \text{vs.} \quad \mathcal{H}_1 : \gamma - 1 > 0$$

A test for the above hypothesis can be based on a consistent estimator of γ.

From (8.84) and (8.85), it follows that, under the assumptions (A) and (B), a consistent estimator of γ is given by

$$\hat{\gamma} = \frac{\hat{Y}_2}{\hat{Y}_1^2} = \frac{n^2}{(n-1)(n+2)} \frac{1}{p} \left[\operatorname{Tr} \mathbf{S}^2 - \frac{1}{n} (\operatorname{Tr} \mathbf{S})^2 \right] / (\operatorname{Tr} \mathbf{S}/p)^2 \tag{8.91}$$

Thus, a test for the sphericity can be based on the statistic

$$T_1 = \hat{\gamma} - 1$$

Under the assumptions (A) and (B), asymptotically

$$\left(\frac{n}{2} \right) (T_1 - \gamma + 1) \sim \mathcal{N} \left(0, \tau^2 \right)$$

where

$$\tau^2 = \frac{2n \left(Y_4 Y_1^2 - 2 Y_1 Y_2 Y_3 + Y_2^3 \right)}{p Y_1^6} + \frac{Y_2^2}{Y_1^4}$$

Under the hypothesis that $\gamma = 1$, and under the assumptions (A) and (B), asymptotically

$$\left(\frac{n}{2} \right) (T_1) \sim \mathcal{N} (0, 1)$$

To evaluate (8.91), we need the following result:

- Assumption 1: $p/n \to c \in (0, \infty)$.
- Assumption 2: $\frac{1}{p} \operatorname{Tr} \mathbf{\Sigma}^k = O(1), k = 1, 2.$
- Assumption 3: $\frac{1}{p} \operatorname{Tr} \mathbf{\Sigma}^k = O(1), k = 3, 4.$

Theorem 8.9.4 (Law of large numbers) Under assumptions 1 through 3, we have

$$\frac{1}{p} \operatorname{Tr} \mathbf{S} \xrightarrow{p} \frac{1}{p} \sum_{i=1}^{p} \lambda_i (\mathbf{\Sigma}) = \alpha$$

$$\frac{1}{p} \operatorname{Tr} \mathbf{S}^2 \xrightarrow{p} (1 + c) \frac{1}{p} \sum_{i=1}^{p} \lambda_i (\mathbf{\Sigma}) + \frac{1}{p} \sum_{i=1}^{p} \left(\lambda_i (\mathbf{\Sigma}) - \alpha \right)^2$$

Theorem 8.9.5 (central limit theorem) Under assumptions 1 and 2, if $\frac{1}{p} \sum_{i=1}^{p} (\lambda_i(\mathbf{\Sigma}) - \alpha)^2 = 0$, then

$$n \times \begin{pmatrix} \frac{1}{p} \operatorname{Tr} \mathbf{S} - \alpha \\ \frac{1}{p} \operatorname{Tr} \mathbf{S}^2 - \frac{n+p+1}{n} \alpha^2 \end{pmatrix} \xrightarrow{d} \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{2}{c} \alpha^2 & 4 \left(1 + \frac{1}{c} \right) \alpha^3 \\ 4 \left(1 + \frac{1}{c} \right) \alpha^3 & 4 \left(\frac{2}{c} + 5 + 2c \right) \alpha^4 \end{pmatrix} \right)$$

where d stands for convergence in distribution.

See [483] for details.

Fisher, Sun and Gallagher [488] modified the tests above. Starting from (8.89), it follows, from the Cauchy–Schwarz inequality, that

$$\left(\sum_{i=1}^{p}\lambda_i^r\right)^2 \leqslant p\left(\sum_{i=1}^{p}\lambda_i^{2r}\right)$$

with equality holding if and only if $\lambda_1 = \cdots = \lambda_p = c$, for all $i = 1, \ldots, p$ and some constant c. Thus, we may consider testing

$$\mathcal{H}_0 : \gamma = 1 \quad \text{vs.} \quad \mathcal{H}_0 : \gamma > 1$$

with

$$\gamma \equiv \frac{\sum_{i=1}^{p}\lambda_i^{2r}/p}{\left(\sum_{i=1}^{p}\lambda_i^r/p\right)^2}$$

We note this test is based on the ratio of arithmetic means of the sample eigenvalues. (8.90) considered in [477] and above where $r = 1$, below we look at the case of $r = 2$.

We make the following assumptions:

- (C) As $p \to \infty$, $Y_i \to Y_i^0$, $0 < Y_i^0 < \infty, i = 1, \ldots, 16$.
- (D) $(n, p) \to \infty, \frac{p}{n} \to c$ where $0 < c < \infty$.

where

$$Y_k = (1/p)\operatorname{Tr}\boldsymbol{\Sigma}^k = \frac{1}{p}\sum_{j=1}^{p}\lambda_j^k(\boldsymbol{\Sigma})$$

An unbiased and (n, p)-consistent estimator of $Y_4 = \frac{1}{p}\sum_{j=1}^{p}\lambda_j^4(\boldsymbol{\Sigma})$ is given by

$$\hat{Y}_4 = \frac{\tau}{p}\left[\operatorname{Tr}\mathbf{S}^4 + b\cdot\operatorname{Tr}\mathbf{S}^3\operatorname{Tr}\mathbf{S} + c_1\cdot\left(\operatorname{Tr}\mathbf{S}^2\right)^2 + d\cdot\operatorname{Tr}\mathbf{S}^2\left(\operatorname{Tr}\mathbf{S}^2\right)^2 + e\cdot\left(\operatorname{Tr}\mathbf{S}^2\right)^4\right] \tag{8.92}$$

where

$$b = -\frac{4}{n}, \quad c_1 = -\frac{2n^2+3n-6}{n\left(n^2+n+2\right)}, \quad d = \frac{2\left(5n+6\right)}{n\left(n^2+n+2\right)}, \quad e = \frac{5n+6}{n\left(n^2+n+2\right)}$$

and

$$\tau = \frac{n^5\left(n^2+n+2\right)}{(n+1)(n+2)(n+4)(n+6)(n-1)(n-2)(n-3)}$$

An unbiased and consistent estimator for Y_2 is given by (8.85). Thus an (n, p)-consistent estimator for the ratio Y_4/Y_2 is provided by

$$\psi = \frac{\hat{Y}_4}{\hat{Y}_2^2}$$

Under assumptions (C) and (D), as $(n,p) \to \infty$, we have

$$\frac{n}{\sqrt{8\left(8+12c+c^2\right)}}\left(\frac{\hat{Y}_4}{\hat{Y}_2^2}-\psi\right) \xrightarrow{d} \mathcal{N}\left(0,\xi^2\right)$$

with

$$\xi^2 = \frac{1}{\left(8+12c+c^2\right)Y_2^6}\Big(\frac{4}{c}Y_4^3 - \frac{8}{c}Y_4Y_2Y_6 - 4Y_4Y_2Y_3^2 + \frac{4}{c}Y_2^2Y_8 + 4Y_6Y_2^3 + 8Y_2^2Y_5Y_3 + 4cY_4Y_2^4 + 8cY_3^2Y_2^3 + c^2Y_2^6\Big) \tag{8.93}$$

Under the null hypothesis, $\psi = 1$, and under the assumptions (C) and (D), as $(n,p) \to \infty$

$$T = \frac{n}{\sqrt{8\left(8+12c+c^2\right)}}\left(\frac{\hat{Y}_4}{\hat{Y}_2^2}-1\right) \xrightarrow{d} \mathcal{N}(0,1) \tag{8.94}$$

For large n and p, the power function of T is

$$\text{Power}_\alpha(T) \approx \Phi\left(\frac{n\left(\frac{Y_4}{Y_2^2}-1\right)}{\xi\sqrt{8\left(8+12c+c^2\right)}} - \frac{z_\alpha}{\xi}\right)$$

under the assumptions (C) and (D), as $(n,p) \to \infty$, we know ξ^2 from (8.93) is constant. From the properties of $\Phi(\cdot)$, it is clear that

$$\text{Power}_\alpha(T) \to 1.$$

Thus the test statistic T in (8.94) is (n,p)-consistent.

Example 8.9.6 (spectrum sensing in cogniive radio) Consider formulating spectrum sensing as a problem of hypothesis testing (see [39]):

$$\mathcal{H}_0 : \quad \mathbf{\Sigma} = \sigma^2\mathbf{I}$$
$$\mathcal{H}_1 : \quad \mathbf{\Sigma} = \mathbf{R}_s + \sigma^2\mathbf{I}$$

where σ^2 is the power of white Gaussian noise (unknown in general) and $\mathbf{R}_s$ is the true covariance matrix of the signal vector. Obviously the above hypothesis testing problem is in the form of (8.88). Many estimators can be used to estimate $\mathbf{R}_s$ which is of low rank. See [40] for details. □

The one-sided tests in this subsection do not need information from the alternative hypothesis. This would be required when using a two-sided test.

8.9.4 Optimal Hypothesis Testing for High-Dimensional Covariance Matrices

Here we use the structure of both $\mathcal{H}_0$ and $\mathcal{H}_1$. Covariance structure plays a fundamental role in multivariate analysis and testing the covariance matrix is an important problem. In a high-dimensional setting, where the dimension p can be comparable to or even much larger than the sample size n, the conventional testing procedures such as the likelihood ratio test (LRT) perform poorly or are not even well defined.

Inspired by applications such as Example 8.9.6, here we consider

$$\mathcal{H}_0 : \mathbf{\Sigma} = \sigma^2 \mathbf{I} \tag{8.95}$$

Cao and Ma [489] investigated this testing problem in high-dimensional settings from a minimax point of view. Consider testing (8.95) against a composite alternative hypothesis

$$\mathcal{H}_1 : \mathbf{\Sigma} \in \Theta, \quad \Theta = \Theta_n = \left\{ \mathbf{\Sigma} : \|\mathbf{\Sigma} - \mathbf{I}\|_F \geqslant \epsilon_n \right\} \tag{8.96}$$

Here, $\|\mathbf{S}\|_F = \left(\sum_{i,j} a_{ij}^2 \right)^{1/2}$ denotes the Frobenius norm of a matrix $\mathbf{A} = (a_{ij})$. It is clear that the difficulty of testing between $\mathcal{H}_0$ and $\mathcal{H}_1$ depends on the value of ϵ_n. The smaller the ϵ_n is, the harder it is to distinguish between the two hypotheses. One naturally asks the following question: What is the boundary that separates the testable region, where it is possible to reliably detect the alternative based on the observations, from the untestable region, where it is impossible to do so? This problem is connected to classical contiguity theory. It is also important to construct a test that can optimally distinguish between the two hypotheses in the testable region. The high-dimensional settings here include all the cases where the dimension $p = p_n \to \infty$ as the sample size $n \to \infty$, and there is no restriction on the limit of p/n unless otherwise stated.

For a given the significance level $0\alpha < 1$, our first goal is to identify the separation rate ϵ_n at which there exists a test ϕ based on the random sample $\left\{ \mathbf{x}_1, \ldots, \mathbf{x}_n \right\}$ such that

$$\inf_{\mathbf{\Sigma} \in \Theta} \mathbb{P}_{\mathbf{\Sigma}} \left(\phi \text{ rejects } \mathcal{H}_0 \right) \geqslant \beta > \alpha$$

Hence the test is able to detect any alternative that is separated away from the null by a certain distance ϵ_n with a guaranteed power $\beta > \alpha$.

The lower and upper bounds together characterize the separation boundary between the testable and nontestable regions when the ratio of the dimension p over the sample size n is bounded. This separation boundary can then be used as a minimax benchmark for the evaluation of the performance of a test in this asymptotic regime.

Lower Bound

We consider the lower bound first. A test $\phi = \phi_n \left(\mathbf{x}_1, \ldots, \mathbf{x}_n \right)$ refers to a measurable function that maps n random vectors to the closed interval $[0, 1]$, where the value stands for the probability of rejecting $\mathcal{H}_0$. So, the significance level of ϕ is

$$\mathbb{P}_I \left(\phi \text{ rejects } \mathcal{H}_0 \right) = \mathbb{E}_I \phi$$

and its power at a certain alternative Σ is

$$\mathbb{P}_{\mathbf{\Sigma}} \left(\phi \text{ rejects } \mathcal{H}_0 \right) = \mathbb{E}_{\mathbf{\Sigma}} \phi$$

Here and after $\mathbb{P}_{\mathbf{\Sigma}}$, $\mathbb{E}_{\mathbf{\Sigma}}$, $\mathrm{Var}_{\mathbf{\Sigma}}$ and $\mathrm{Cov}_{\mathbf{\Sigma}}$ denote the induced probability measure, expectation, variance and covariance when

$$\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{i.i.d}{\sim} \mathcal{N}_p (0, \mathbf{\Sigma})$$

Let $\epsilon_n = b\sqrt{p/n}$ for some constant b, and define

$$\Theta (b) = \left\{ \mathbf{\Sigma} : \|\mathbf{\Sigma} - \mathbf{I}\|_F \geqslant b\sqrt{p/n} \right\} \tag{8.97}$$

Theorem 8.9.7 (lower bound) Let $0 < \alpha < \beta < 1$. Suppose that $n \to \infty$, $p \to \infty$, and that $p/n \leq \kappa$ for some constant $\kappa < \infty$ and all n. Then there exists a constant $b = b(\kappa, \beta - \alpha) < 1$, such that for any test ϕ with significance level α for testing $\mathcal{H}_0 : \mathbf{\Sigma} = \mathbf{I}$,

$$\limsup_{n\to\infty} \inf_{\mathbf{\Sigma}\in\Theta(b)} \mathbb{E}_{\mathbf{\Sigma}}\phi < \beta$$

Upper Bound

There is a level α whose power over Θ_n is uniformly larger than a prescribed value $\beta > \alpha$, if $_n = b\sqrt{p/n}$ for a large-enough constant b. This is in agreement with the lower bound result in Theorem 8.9.7 when p/n is bounded.

In addition, the results in the current section remain valid even when p/n is unbounded. This is the ultra-high-dimensional setting: both $n, p \to \infty$ and $p/n \to \infty$. The LRT and corrected LRT [160] are not well defined in this case. The testing problem in this asymptotic regime is not as well studied as in the previous categories. Birke and Dette [490] derived the asymptotic null distribution of the Ledoit–Wolf test under the current asymptotic regime. More recently, Chen *et al.* [476] proposed a new test statistic and derived its asymptotic null distribution when both both $n, p \to \infty$, regardless of the limiting behavior of p/n.

Let us first start with test statistic. Given a random i.i.d. sample $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$, a natural approach to test between (8.95) and (8.96) is to first estimate the squared Frobenius norm

$$\|\mathbf{\Sigma} - \mathbf{I}\|_F^2 = \mathrm{Tr}\,(\mathbf{\Sigma} - \mathbf{I})^2$$

by some statistic $T_n = T_n\left(\mathbf{x}_1, \ldots, \mathbf{x}_n\right)$, and then reject the null hypothesis $\mathcal{H}_0$ if T_n is too large. To estimate $\|\mathbf{\Sigma} - \mathbf{I}\|_F^2 = \mathrm{Tr}\,(\mathbf{\Sigma} - \mathbf{I})^2$, note that

$$\mathbb{E}_{\mathbf{\Sigma}} d\left(\mathbf{x}_1, \mathbf{x}_2\right) = \mathrm{Tr}\,(\mathbf{\Sigma} - \mathbf{I})^2$$

where

$$d\left(\mathbf{x}_1, \mathbf{x}_2\right) = \left(\mathbf{x}_1^H \mathbf{x}_2\right)^2 - \left(\mathbf{x}_1^H \mathbf{x}_1 + \mathbf{x}_2^H \mathbf{x}_2\right) + p \tag{8.98}$$

Therefore, $\mathrm{Tr}\,(\mathbf{\Sigma} - \mathbf{I})^2$ can be estimated by the following U-statistic

$$T_n = \frac{2}{n(n-1)} \sum_{1\leqslant i<j\leqslant n} d\left(\mathbf{x}_i, \mathbf{x}_j\right) \tag{8.99}$$

for which we have

$$\mathrm{Var}_n(\mathbf{\Sigma}) = \mathrm{Var}_{\mathbf{\Sigma}}\left(T_n\right) = \frac{4}{n(n-1)}\left[\left(\mathrm{Tr}\left(\mathbf{\Sigma}^2\right)\right)^2 + \mathrm{Tr}\left(\mathbf{\Sigma}^4\right)\right] + \frac{8}{n}\,\mathrm{Tr}\left(\mathbf{\Sigma}^2(\mathbf{\Sigma} - \mathbf{I})^2\right)$$

Proposition 8.9.8 (Theorem 2 of [476]) Suppose that $p \to \infty$ as $n \to \infty$. If a sequence of covariance matrices satisfy $\mathrm{Tr}\left(\mathbf{\Sigma}^2\right) \to \infty$ and $\mathrm{Tr}\left(\mathbf{\Sigma}^4\right)/\mathrm{Tr}^2\left(\mathbf{\Sigma}^2\right) \to 0$ as $n \to \infty$. then under $\mathbb{P}_{\mathbf{\Sigma}}$, we have

$$\frac{T_n - \mu_n(\mathbf{\Sigma})}{\sigma_n(\mathbf{\Sigma})} \Rightarrow \mathcal{N}(0, 1)$$

Note that as $n \to \infty$, the identity matrix $\mathbf{I}_{p\times p}$ satisfies the condition of the above proposition. Also note that $\mu_n(\mathbf{I}) = 0$, and $\sigma_n^2(\mathbf{\Sigma}) = \frac{4p(p-1)}{n(n-1)}$. Proposition 8.9.8 describes

the behavior of test statistic T_n under $\mathcal{H}_0$, and we could define the test as the following. For any $\alpha \in (0,1)$, an asymptotic level α test based on T_n is given by

$$\psi = I\left(T_n > z_{1-\alpha} \cdot 2\sqrt{\frac{p(p-1)}{n(n-1)}}\right) \tag{8.100}$$

Here, $I(\cdot)$ is the indicator function, and $z_{1-\alpha}$ denotes the $100 \times (1-\alpha)$-th percentile of the standard normal distribution.

We now study the rate of convergence for the distribution of $\left[T_n - \mu_n(\mathbf{\Sigma})\right]/\sigma_n(\mathbf{\Sigma})$ to its normal limit in Kolmogorov distance. Let $\Phi(\cdot)$ be the cumulative distribution function of the standard normal distribution. We have the following Berry–Essen type bound.

Proposition 8.9.9 Under the condition of Proposition 8.9.8, there exists a numeric constant C such that

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}_{\mathbf{\Sigma}}\left(\frac{T_n - \mu_n(\mathbf{\Sigma})}{\sigma_n(\mathbf{\Sigma})} \leqslant x\right) - \Phi(x)\right| \leqslant C\left[\frac{1}{n} + \frac{\mathrm{Tr}\left(\mathbf{\Sigma}^4\right)}{\mathrm{Tr}^2\left(\mathbf{\Sigma}^2\right)}\right]^{1/5}$$

Equipped with Proposition 8.9.9, we now investigate the power of the test (8.100) over the composite alternative $\mathcal{H}_1 : \mathbf{\Sigma} \in \Theta(b)$, with $b < 1$, $\Theta(b)$ is defined in (8.97).

Theorem 8.9.10 (upper bound) Suppose that $p \to \infty$ as $n \to \infty$. For any significance level $\alpha \in (0,1)$, and $\Theta(b)$ defined in (8.97), the power of the test in (8.100) satisfies

$$\lim_{n\to\infty}\inf_{\mathbf{\Sigma}\in\Theta(b)} \mathbb{E}_{\mathbf{\Sigma}}\psi = 1 - \Phi\left(z_{1-\alpha} - \frac{b^2}{2}\right) > \alpha$$

Moreover, for $b_n \to \infty$, $\lim_{n\to\infty}\inf_{\mathbf{\Sigma}\in\Theta(b_n)} \mathbb{E}_{\mathbf{\Sigma}}\psi = 1$

In the classical asymptotic regime where p is fixed and $n \to \infty$, the likelihood ratio test (LRT) is one of the most commonly used tests. In the high-dimensional setting where both n and p are large and $p < n$, Bai *et al.* [160] showed that the LRT is not well behaved as the chi-squared limiting distribution under $\mathcal{H}_0$ no longer holds.

When p fixed and $n \to \infty$. In this classical asymptotic regime, conventional tests for (8.95) include the likelihood ratio test (LRT) [371], Roy's largest root test [491], and Nagao's test [492]. In particular, the LRT statistic is $\mathrm{LR}_n = nL_n$, where

$$L_n = \mathrm{Tr}\,\mathbf{S} - \log\det(\mathbf{S}) - p$$

The asymptotic distribution of $\mathrm{LR}_n = nL_n$ under $\mathcal{H}_0$ is chi-square distributed $\chi^2_{p(p+1)/2}$.

For testing (8.95), when $p < n$ and $p/n \to c_n \in (0,1)$, Bai *et al.* [160] proposed a corrected likelihood ratio test (CLRT) with the test statistic CLR_n given by

$$\mathrm{CLR}_n = \frac{L_n - p\left[1 - \left(1 - c_n^{-1}\right)\log\left(1 - c_n\right)\right] - \frac{1}{2}\log\left(1 - c_n\right)}{\sqrt{-2\log\left(1 - c_n\right) - 2c_n}} \tag{8.101}$$

whose asymptotic null distribution is $\mathcal{N}(0,1)$. No test based on the likelihood ratio can be defined when $p > n$ or $c_n > 1$. The power of the CLRT is uniformly dominated by that of given in (8.100) over the entire asymptotic regime under which the CLRT is applicable.

We have focused in this section on testing the hypotheses under the Frobenius norm. The technical arguments developed here can also be used for testing under other matrix norms such as the spectral norm.

8.9.5 Sphericity Test

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be i.i.d. $\mathbb{R}^p$-valued random vectors from a normal distribution $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} \in \mathbb{R}^p$ is the mean vector and $\boldsymbol{\Sigma}$ the $p \times p$ covariance matrix. Consider the hypothesis test:

$$\mathcal{H}_0 : \boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_p \quad \text{vs.} \quad \mathcal{H}_1 : \boldsymbol{\Sigma} \neq \sigma^2 \mathbf{I}_p \tag{8.102}$$

where σ^2 is unknown. The identity hypothesis in (8.102) covers the hypothesis for

$$\mathcal{H}_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 \quad \text{vs.} \quad \mathcal{H}_1 : \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}_0$$

for an *arbitrary* specifically known invertible covariance matrix $\boldsymbol{\Sigma}_0$. Thus the assumption of (8.102) will not lose generality. Denote

$$\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i, \quad \mathbf{A} = \sum_{i=1}^{n} \left(\mathbf{x}_i - \overline{\mathbf{x}}\right)\left(\mathbf{x}_i - \overline{\mathbf{x}}\right)', \quad \text{and} \quad \mathbf{S} = \frac{1}{n-1}\mathbf{A} \tag{8.103}$$

The likelihood ratio statistic of test (8.102) was first derived by Mauchly [493] as

$$V_n = \det \mathbf{A} \cdot \left(\frac{1}{p} \operatorname{Tr} \mathbf{A}\right)^{-p} = \det \mathbf{S} \cdot \left(\frac{1}{p} \operatorname{Tr} \mathbf{S}\right)^{-p} \tag{8.104}$$

Note that the matrices $\mathbf{A}, \mathbf{S}$ are not of full rank when $p > n$ and consequently their determinants are equal to zero in this case. This says that the likelihood ratio test of (8.102) only exists for $p \leq n$. The statistic V_n is commonly known as the ellipticity statistic.

By Theorem 3.1.2 and Cololary 3.2.19 from Muirhead [37], under $\mathcal{H}_0$ in (8.102)

$$\frac{n}{\sigma^2} \cdot \mathbf{S} \quad \text{and} \quad \mathbf{Z}'\mathbf{Z} \ \text{ have the same distribution} \tag{8.105}$$

where $\mathbf{Z} = \left\{z_{ij}\right\}_{(n)\times p}$ and $z'_{ij}s$ are i.i.d. with distribution with $\mathcal{N}(0,1)$. This says that, with probability one, $\mathbf{S}$ is not of full rank when $p \geq n$, and consequently $\det \mathbf{A} = 0$. This indicates that the likelihood ratio test of (8.102) only exists when $p < n$.

Gleser [494] proved the likelihood ratio test with the rejection region $\left\{V_n \leqslant c_\alpha\right\}$ (where c_α is chosen so that the test has a significant level of α) is unbiased. The distribution of the test statistic V_n can be studied through its moments. When the null hypothesis $\mathcal{H}_0 : \boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_p$ is true, the following result is cited from p. 341 of Muirhead [37]:

$$\mathbb{E}\left(V_n^h\right) = p^{ph} \frac{\Gamma\left[\frac{1}{2}(n-1)p\right]}{\Gamma\left[\frac{1}{2}(n-1)+ph\right]} \frac{\Gamma\left[\frac{1}{2}(n-1)+h\right]}{\Gamma_p\left[\frac{1}{2}(n-1)\right]} \quad \text{for} \quad h > -\frac{1}{2} \tag{8.106}$$

When p is assumed a fixed integer, the following result, referenced from section 10.7.4 of of Muirhead [37] and and section 8.3.3 of Anderson [495], gives an explicit expansion

of the distribution function of $-2\rho \log\left(V_n\right)$, $\rho = 1 - \left(2p^2 + p + 2\right)/\left(6np - 6p\right)$, as $M = \rho\,(n-1) \to \infty$:

$$\begin{aligned}&\mathbb{P}\left(-(n-1)\,\rho \log\left(V_n\right) \leqslant x\right)\\ &\quad = \mathbb{P}\left(\chi_f^2 \leqslant x\right) + \frac{\gamma}{M^2}\left[\mathbb{P}\left(\chi_{f+4}^2 \leqslant x\right) - \mathbb{P}\left(\chi_f^2 \leqslant x\right)\right] + O\left(M^{-3}\right)\end{aligned} \tag{8.107}$$

where $f = (p+2)\,(p-1)\,/2$, $\gamma = (n-1)^2\rho^2\omega_2$, and ω_2 given by

$$\omega_2 = \frac{(p-1)\,(p-2)\,(p+2)\left(2p^2+6p^2+3p+2\right)}{288p^2(n-1)^2\rho^2} \tag{8.108}$$

In other words, this classical asymptotic result shows that

$$-n\rho \log V_n \quad \text{converges to} \quad \chi^2(f) \tag{8.109}$$

in distribution as $n \to \infty$ with p fixed. The quantity ρ is a correction term to improve the convergence rate.

Nagarsenker and Pillai [496] tabulated the lower 5 percentile and 1 percentile of the asymptotic distribution of V_n under the null hypothesis $\mathcal{H}_0 : \mathbf{\Sigma} = \sigma^2\mathbf{I}_p$.

A different test for sphericity other than the likelihood ratio test was recommended by John [497] who studied the test statistic

$$U = \frac{1}{p}\left[\frac{\mathbf{S}}{(1/p\,\mathrm{Tr}\,\mathbf{S})} - \mathbf{I}_p\right]^2 = \frac{(1/p)\,\mathrm{Tr}\,\mathbf{S}^2}{(1/p\,\mathrm{Tr}\,\mathbf{S})^2} - 1 \tag{8.110}$$

It follows from John [497] that the test that rejects the null hypothesis when $U > c_\alpha$, where c_α is determined by the significant level of α, is a locally most powerful invariant test for sphericity and this test is more general than the aforementioned likelihood ratio test because it can be performed even with $p > n$ when the sample size is smaller than the dimension. John [498] further showed that under the null hypothesis of (8.102), the limiting distribution of the test statistic U, as the sample size n goes to infinity while the dimension p remains fixed, is given by

$$\frac{1}{2}npU \xrightarrow{d} \chi^2_{p(p+1)/2-1} \tag{8.111}$$

Finally, when $p \geq n$, we know the LRT does not exist as mentioned below. There are some recent works on choosing other statistics to study the spherical test of (8.102). Along with this line, Ledoit and Wolf [469] re-examined the limiting distribution of the test statistic U in the high-dimensional situation where $\lim\limits_{n\to\infty} \frac{p}{n} = c \in (0, \infty)$. They proved that, under the null hypothesis of (8.102)

$$nU_n - p \xrightarrow{d} \mathcal{N}(1, 4) \tag{8.112}$$

Ledoit and Wolf further argued that since

$$\frac{2}{p}\chi^2_{p(p+1)/2-1} - p \xrightarrow{d} \mathcal{N}(1, 4) \tag{8.113}$$

John's n-asymptotic results (assuming p is fixed) of test statistic U still remains valid in the high-dimensional case (i.e. both p and n are large). Chen, Zhang and Zhong [476] extended Ledoit and Wolf's asymptotic result to non-normal distributions with certain conditions on their covariance matrices. Cai and Ma [484] considered

testing a covariance matrix $\mathbf{\Sigma}$ in a high-dimensional setting where the dimension p can be comparable to, or much larger than, the sample size n. The problem of testing the hypothesis $\mathcal{H}_0 : \mathbf{\Sigma} = \mathbf{\Sigma}_0$ for a given covariance matrix $\mathbf{\Sigma}_0$ is studied from a minimax point of view. A test based on a U-statistic is introduced and is shown to be rate optimal over this asymptotic regime. It is shown that the power of this test uniformly dominates that of the corrected likelihood ratio test (CLRT) (first proposed by Bai *et al.* [160]) over the entire asymptotic regime under which the CLRT is applicable.

Here we focus on the likelihood ratio test for sphericity in the high dimensional case $\lim_{n\to\infty} \frac{p}{n} = y \in (0, 1)$, due to [499], and develop a central limit theorem for the likelihood ratio test statistic $\log V_n$ as given in (8.104).

Theorem 8.9.11 Assume that $p := p_n$ is a series of positive integers depending on n such that $n > 1 + p$ for all $n \geq 3$ and $p/n \to y \in (0, 1]$ as $n \to \infty$. Let V_n be defined as given in (3.3). Then under $\mathcal{H}_0 : \mathbf{\Sigma} = a^2\mathbf{I}_p$ (a^2 unknown), $\left(\log V_n - \mu_n\right)/\sigma_n$ converges in distribution to $\mathcal{N}(0, 1)$ as $n \to \infty$, where

$$\mu_n = -p - (n - p - 1.5)\log\left(1 - \frac{p}{n-1}\right)$$
$$\sigma_n^2 = -2\left[\frac{p}{n-1} + \log\left(1 - \frac{p}{n-1}\right)\right] > 0$$

Though a^2 is unspecified, the limiting distribution in Theorem 8.9.11 is pivotal, that is, it does not depend on a^2. This is because a^2 is canceled in the expression of (8.104): $\det(\alpha\mathbf{S}) = \alpha^p \det(\mathbf{S})$ and $(\mathrm{Tr}(\alpha\mathbf{S}))^{-p} = \alpha^{-p} \cdot (\mathrm{Tr}(\mathbf{S}))^{-p}$ for any $\alpha > 0$.

Simulation by Yang [499] indicates that, in regard to the test size (or alpha error), the proposed high-dimensional LRT using Theorem 8.9.11 shows noninferiority to the traditional LRT when p is small, yet a significant improvement over the traditional one when p becomes large.

Traditionally, the likelihood ratio tests (LRT) for the mean vectors and covariance matrices of normal distributions were performed by using the chi-square approximation to the limiting distributions of the likelihood ratio test statistics. However, this approximation relies on a theoretical assumption that the sample size n goes to infinity, while the dimension p remains fixed. In practice, this requires the dataset to have a large sample size n but a low dimension p. Simulation in [474] shows that the chi-square approximation in (8.109) is far from reasonable when p is large. As many modern datasets feature high dimensions, these traditional likelihood ratio tests were shown to be less accurate in analyzing those datasets.

On the sphericity test with large-dimensional observations see [500].

8.9.6 Testing Equality of Multiple Covariance Matrices of Normal Distributions

Suppose that the sample covariance matrices $\mathbf{S}_1, \ldots, \mathbf{S}_k$ have been computed from independent random samples from multivariate normal distributions with covariance matrices $\mathbf{\Sigma}_1, ..., \mathbf{\Sigma}_k$, respectively. We test whether the hypothesis

$$\mathcal{H}_0 : \mathbf{\Sigma}_1 = \cdots = \mathbf{\Sigma}_k$$

is true. The alternative hypothesis $\mathcal{H}_1$ is that $\mathcal{H}_0$ is not true.

Let $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i}$ be i.i.d. $\mathbb{R}^p$-valued random vectors from k p-variate normal distributions $\mathcal{N}_p\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)$ for $i = 1, \ldots, k$, where $k \geq 2$ is a fixed integer. Consider the hypothesis test that these k normal distributions have a common but unknown covariance matrix:

$$\mathcal{H}_0 : \boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_k \tag{8.114}$$

Denote

$$\overline{\mathbf{x}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x}_{ij}, \quad \mathbf{A}_i = \sum_{j=1}^{n_i} \left(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i\right) \left(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i\right)', \quad \text{for} \quad i = 1, \ldots, k$$

and

$$\mathbf{A} = \mathbf{A}_i + \cdots \mathbf{A}_k, \quad n = n_1 + \cdots + n_k \tag{8.115}$$

Wilks [501] gave the likelihood ratio test of (8.114) with a test statistic

$$\Lambda_n = \frac{\prod_{i=1}^{k} \left(\det \mathbf{A}_i\right)^{n_i/2}}{(\det \mathbf{A})^{n/2}} \cdot \frac{n^{np/2}}{\prod_{i=1}^{k} n_i^{n_i p/2}} \tag{8.116}$$

and the test rejects the null hypothesis $\mathcal{H}_0$ at $\Lambda_n \leqslant c_\alpha$, where the critical value c_α is determined so that the test has the significant level of α.

Any of the matrices $\mathbf{A}_i, i = 1, 2, \ldots, k$ will not have a full rank when $p > n_i$ for any $i = 1, \ldots, k$ and consequently its determinants are equal to zero, so are the test statistic Λ_n. Therefore, the likelihood ratio test of (8.114) only exists when $p \leq n_i$ for all $i = 1, \ldots, k$. Another drawback of the likelihood ratio test is on its bias (see section 8.2.2 of [37]). Bartlett [502] suggested using a modiffied likelihood ratio test statistic Λ_n^* by substituting every sample size n_i with its degree of freedom $n_i - 1$ and substituting the total sample size n with $n - k$:

$$\Lambda_n^* = \frac{\prod_{i=1}^{k} \left(\det \mathbf{A}_i\right)^{(n_i-1)/2}}{\left(\det \mathbf{A}_i\right)^{(n-k)/2}} \cdot \frac{(n-k)^{(n-k)p/2}}{\prod_{i=1}^{k} \left(n_i - 1\right)^{(n_i-1)p/2}} \tag{8.117}$$

The unbiased property of this modified likelihood ratio test was proved by Sugiura and Nagao [503] for $k = 2$ and by Perlman [504] for a general k.

Many traditional tests assume a fixed p. Recently, Schott [505] studied an alternative test of (8.114) based on the Wald statistic

$$T = \frac{n}{2} \sum_{i \neq j}^{k} \operatorname{Tr}\left\{ \left(\frac{1}{n_i} \mathbf{S}_i - \frac{1}{n_j} \mathbf{S}_j \right) \left(\frac{1}{n} \mathbf{S} \right)^{-1} \right\}^2 \tag{8.118}$$

where the Wald statistic is well defined as long as $\mathbf{S}$ is nonsingular, hence it only requires $n = n_1 + \cdots + n_k \geqslant p$. Schott [505] showed that if p remains fixed, then the limiting null distribution of T as $n_i \to \infty$ for $i = 1, \ldots, k$, is a chi-square distribution with degree of

freedom $(k-1)p(p+1)/2$. For the high-dimensional settings, Schott [481] proposed a modified Wald statistic

$$T_{np} = \sum_{i<j} \left\{ \operatorname{Tr}\left(\mathbf{S}_i - \mathbf{S}_j\right)^2 - \frac{1}{n_i \eta_i}\left[n_i\left(n_i-2\right)\operatorname{Tr}\mathbf{S}_i^2 + n_i^2(\operatorname{Tr}\mathbf{S})^2\right] \right.$$
$$\left. - \frac{1}{n_j \eta_j}\left[n_j\left(n_j-2\right)\operatorname{Tr}\mathbf{S}_j^2 + n_j^2(\operatorname{Tr}\mathbf{S})^2\right] \right\} \tag{8.119}$$

with $\eta_i = \left(n_i+2\right)\left(n_i-1\right)$ and showed that this statistic T_{np} is an unbiased estimator for $\sum_{i\neq j} \operatorname{Tr}\left(\mathbf{\Sigma}_i - \mathbf{\Sigma}_j\right)^2$. Schott further proved that assuming

- (1) $p_i/n \to y_i \in [0,\infty)$, and at least one $y_i > 0$;
- (2) $\lim_{n\to\infty} \frac{1}{p} \operatorname{Tr}\mathbf{\Sigma}^k = \gamma_k \in [0,\infty)$, for $k = 1,\ldots,8$

this statistic T_{np} as defined in (8.119) converges in distribution to $\mathcal{N}(0,\sigma^2)$, where

$$\sigma^2 = 4\left[\sum_{i\neq j}\left(y_i + y_j\right)^2 + (k-1)(k-2)\sum_{i=1}^{k} y_i^2\right]\gamma_2^2$$

Here we develop the likelihood ratio test for testing equality of multiple covariance matrices of normal distributions in the high-dimensional settings $\frac{p}{n} \to y \in (0,1)$, as $n,p \to \infty$. Our proposed test is based on the following central limit theorem due to [499] for the likelihood ratio statistic $\log\Lambda_n$ under the null hypothesis (8.114).

Theorem 8.9.12 Assume $n_i = n_i(p)$ for all $1 \le i \le k$ such that $\min_{1\leqslant i\leqslant k} n_i > p+1$ and $\lim_{p\to\infty} p/n_i = y_i \in (0,1]$ as $p \to \infty$ for each $1 \le i \le k$. Let $n = n_1 + \cdots + n_k$ and Λ_n^* be defined as in (8.117). Then under

$$\mathcal{H}_0 : \mathbf{\Sigma}_1 = \cdots = \mathbf{\Sigma}_k$$

the sequence

$$\frac{\log\Lambda_n^* - \mu_n}{(n-k)\,\sigma_n}$$

converges in distribution to $\mathcal{N}(0,1)$ as $p \to \infty$, where

$$\mu_n = \frac{1}{2}\left[(n-k)(n-k-0.5)\log\left(1-\frac{p}{n-k}\right)\right.$$
$$\left. - \sum_{i=1}^{k}\left(n_i-1\right)\left(n_i-p-1.5\right)\log\left(1-\frac{p}{n_i-k}\right)\right]$$
$$\sigma_n^2 = \frac{1}{2}\left[\log\left(1-\frac{p}{n-k}\right) - \sum_{i=1}^{k}\left(\frac{n_i-1}{n-k}\right)^2\log\left(1-\frac{p}{n_i-k}\right)\right] > 0$$

8.9.7 Testing Independence of Components of Normal Distribution

For a multivariate distribution $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we partition a set of p variates with a joint normal distribution into k subsets and ask whether the k subsets are mutually independent, or equivalently, we want to test whether variables among different subsets are dependent. In particular, let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be i.i.d. $\mathbb{R}^p$-valued random vectors with normal distribution $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Let the p-component vector $\mathbf{x}$ be distributed according to $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We partition $\mathbf{x}$ into k subvectors:

$$\mathbf{x} = \left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(k)}\right)' \tag{8.120}$$

where each $\mathbf{x}^{(i)}$ has dimension p_i, respectively, with $p = \sum_{i=1}^{k} p_i$. The vector of means $\boldsymbol{\mu}$ and the covariance matrices $\boldsymbol{\Sigma}$ are partitioned similarly:

$$\boldsymbol{\mu} = \left(\boldsymbol{\mu}^{(1)}, \ldots, \boldsymbol{\mu}^{(k)}\right)' \tag{8.121}$$

and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \cdots & \boldsymbol{\Sigma}_{1k} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} & \cdots & \boldsymbol{\Sigma}_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{k1} & \boldsymbol{\Sigma}_{k2} & \cdots & \boldsymbol{\Sigma}_{kk} \end{pmatrix}$$

The null hypothesis is that the subvectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(k)}$ are mutually independently distributed: the density of $\mathbf{x}$ factors into the product of the density functions of $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(k)}$:

$$\mathcal{H}_0 : f(\mathbf{x} \,|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{k} f\left(\mathbf{x}^{(i)} \,\middle|\, \boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma}_{ii}\right) \tag{8.122}$$

If $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(k)}$ are independent subvectors, then the covariance matrix is block diagonal and denoted by $\boldsymbol{\Sigma}_0$.

The block diagonal covariance matrix $\boldsymbol{\Sigma}_0$ is written as

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & 0 & \cdots & 0 \\ 0 & \boldsymbol{\Sigma}_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{\Sigma}_{kk} \end{pmatrix}$$

with $\boldsymbol{\Sigma}_{ii}$ unspecified for $1 \le i \le k$. Given a sample of size n, $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are n observations on random vector $\mathbf{x}$, the likelihood ratio is

$$\Lambda_n = \frac{\max\limits_{\boldsymbol{\mu}, \boldsymbol{\Sigma}_0} L\left(\mu, \boldsymbol{\Sigma}_0\right)}{\max\limits_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} L\left(\mu, \boldsymbol{\Sigma}\right)} \tag{8.123}$$

where the likelihood function is

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{k} \frac{1}{(2\pi)^{p/2} \det^{1/2}(\boldsymbol{\Sigma})} \exp\left\{-\frac{1}{2}\left(\mathbf{x}_i - \boldsymbol{\mu}\right)' \boldsymbol{\Sigma}\left(\mathbf{x}_i - \boldsymbol{\mu}\right)\right\} \tag{8.124}$$

$L\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}_0\right)$ is $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\{\boldsymbol{\Sigma}_{ij} = 0,\ i \ne j$, for all $0 \le i, j \le k$; and the maximum is taken with respect to all vectors $\boldsymbol{\mu}$ and positive definite matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_0$. According to Theorem 11.2.2 of Muirhead [37], we have

$$\max_{\boldsymbol{\mu},\boldsymbol{\Sigma}} L(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{pn/2}\det^{n/2}\left(\hat{\boldsymbol{\Sigma}}_{\Omega}\right)} \exp\left\{-\frac{1}{2}pn\right\} \tag{8.125}$$

where

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{x}_i - \overline{\mathbf{x}}\right)'\left(\mathbf{x}_i - \overline{\mathbf{x}}\right) \tag{8.126}$$

Under the null hypothesis,

$$\begin{aligned}\max_{\boldsymbol{\mu},\boldsymbol{\Sigma}_0} L\left(\boldsymbol{\mu},\boldsymbol{\Sigma}_0\right) &= \prod_{i=1}^{k} \max L_i\left(\boldsymbol{\mu}^{(i)},\boldsymbol{\Sigma}_{ii}\right) \\ &= \prod_{i=1}^{k} \frac{1}{(2\pi)^{p_i n/2}\det^{n/2}\left(\hat{\boldsymbol{\Sigma}}_{ii}\right)} \exp\left\{-\frac{1}{2}p_i n\right\} \\ &= \frac{1}{(2\pi)^{pn/2}\prod_{i=1}^{k}\det^{n/2}\left(\hat{\boldsymbol{\Sigma}}_{ii}\right)} \exp\left\{-\frac{1}{2}pn\right\}\end{aligned}$$

where

$$\hat{\boldsymbol{\Sigma}}_{ii} = \frac{1}{n-1}\sum_{j=1}^{n}\left(\mathbf{x}_j^{(i)} - \overline{\mathbf{x}}^{(i)}\right)\left(\mathbf{x}_j^{(i)} - \overline{\mathbf{x}}^{(i)}\right)' \tag{8.127}$$

If we partition $\mathbf{S}$ and $\hat{\boldsymbol{\Sigma}}$ in the same way as $\boldsymbol{\Sigma}$, we find that

$$\hat{\boldsymbol{\Sigma}}_{ii} = \frac{1}{n-1}\mathbf{S}_{ii}$$

Then the likelihood ratio becomes

$$\Lambda_n = \frac{\max_{\boldsymbol{\mu},\boldsymbol{\Sigma}_0} L\left(\boldsymbol{\mu},\boldsymbol{\Sigma}_0\right)}{\max_{\mu,\boldsymbol{\Sigma}} L(\mu,\boldsymbol{\Sigma})} = \frac{\det^{n/2}\left(\hat{\boldsymbol{\Sigma}}_{\Omega}\right)}{\prod_{i=1}^{k}\det^{n/2}\left(\hat{\boldsymbol{\Sigma}}_{ii}\right)} = \frac{\det^{n/2}(\mathbf{S})}{\prod_{i=1}^{k}\det^{n/2}\left(\mathbf{S}_{ii}\right)} \tag{8.128}$$

The critical region of the likelihood ratio test is

$$\Lambda_n \leqslant \Lambda_n(\alpha) \tag{8.129}$$

where $\Lambda_n(\alpha)$ is a number such that the probability of (8.129) is α when $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$.

Wilks' Statistic

We want to test the hypothesis

$$\mathcal{H}_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 \quad \text{vs.} \quad \mathcal{H}_1 : \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}_0 \tag{8.130}$$

Now we employ Wilks' statistic to do the test. Let

$$W_n = \frac{\det(\mathbf{S})}{\prod_{i=1}^{k}\det\left(\mathbf{S}_{ii}\right)} \tag{8.131}$$

W_n can be expressed entirely in terms of sample correlation coefficients. $\Lambda_n = W_n^{n/2}$ is a monotonically increasing function of W_n. The critical region can be equivalently written as $W_n \leqslant W_n(\alpha)$. $W_n = 0$ if $p > n$, since the matrix $\mathbf{S}$ is not of full rank in this case. Set

$$f = \frac{1}{2}\left(p^2 - \sum_{i=1}^{k} p_i^2\right) \quad \text{and} \quad \rho = 1 - \frac{2\left(p^3 - \sum_{i=1}^{k} p_i^3\right) + 9\left(p^2 - \sum_{i=1}^{k} p_i^2\right)}{6(n+1)\left(p^2 - \sum_{i=1}^{k} p_i^2\right)} \tag{8.132}$$

When $n \to \infty$ while all $p_i's$ remain fixed, the traditional χ^2 approximation to the distribution of Λ_n is found in Theorem 11.2.5 in Muirhead [37]:

$$-2\rho \log\left(\Lambda_n\right) \xrightarrow{d} \chi_f^2$$

When p is large enough or is proportional to n, this chi-square approximation may fail [474]. In fact, their results show that the central limit theorem (CLT) holds: $\left(\log W_n - \mu_n\right)/\sigma_n$ actually converges to the standard normal for a fixed number of partition k, where μ_n and σ_n can be expressed explicitly as a function of sample size and partition.

Considering the insufficiency of the LRT when p is large, Bai *et al.* [160] suggested a so-called corrected likelihood ratio test for covariance matrices of Gaussian populations when the dimension is large compared to the sample size. They also used a LRT to fit high-dimensional normal distribution $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\mathcal{H}_0 : \boldsymbol{\Sigma} = \mathbf{I}_p$. In their derivation, the dimension p is no longer a fixed constant, but rather is a variable that goes to infinity along with the sample size n, and the ratio between $p = p_n$, and converges to a constant y:

$$\lim_{n\to\infty} \frac{p}{n} = y \in (0, 1)$$

Jiang *et al.* [473] further extend Bai's result to cover the case of $y = 1$, and reached the CLT of the LRT used for testing dependence of k groups of components for high-dimensional datasets, where k is a fixed number. Jiang and Yang [474] studied several other classical likelihood ratio tests for means and covariance matrices of high-dimensional normal distributions. Most of these tests have the asymptotic results for their test statistics derived decades ago under the assumption of a large n but a fixed p. Their results supplement these traditional results in providing alternatives to analyze high-dimensional datasets including the critical case $p/n \to 1$.

Zhang [506] has proven the CLT for the LRT, allowing that k changes with n and the partition can be unbalanced in the sense that numbers of components within subsets are not necessarily proportional. The main result is summarized below.

Let $k \geq 2$, $p_1, \ldots, p_k$ be an arbitrary partition of dimension p. Denote $p = \sum_{k=1}^{k} p_i$, and let

$$\boldsymbol{\Sigma} = \left(\boldsymbol{\Sigma}_{ij}\right)_{p\times p} \tag{8.133}$$

be the covariance matrix (positive definite), where $\boldsymbol{\Sigma}_{ij}$ is a $p_i \times p_i$ sub-matrix for all $1 \leq i, j \leq k$. Consider the following hypothesis

$$\mathcal{H}_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 \quad \text{vs.} \quad \mathcal{H}_1 : \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}_0 \tag{8.134}$$

which is equivalent to hypothesis (8.122). Let $\mathbf{S}$ be the sample covariance matrix. Then partition $\mathbf{A} = (n-1)\mathbf{S}$ in the following way:

$$\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \cdots & \mathbf{A}_{1k} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \cdots & \mathbf{A}_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_{k1} & \mathbf{A}_{k2} & \cdots & \mathbf{A}_{kk} \end{pmatrix}$$

where $\mathbf{A}_{ij}$ is a $p_i \times p_j$ matrix. Wilks [501] suggested the likelihood ratio statistic for test (8.122)

$$\Lambda_n = \frac{\det^{n/2}(\mathbf{A})}{\prod_{i=1}^{k} \det^{n/2}(\mathbf{A}_{ii})} = \left(W_n\right)^{n/2} \tag{8.135}$$

When $p > n+1$, the matrix $\mathbf{A}$ is not full rank, and thus Λ_n is degenerate. We have the following theorem due to [506].

Theorem 8.9.13 Let p satisfy $p < n-1$ and $p \to \infty$ as $n \to \infty$. $p_1, \ldots, p_k$ are k integers such that $p = \sum_{i=1}^{k} p_i$ and $\frac{\max_i p_i}{p} \leqslant 1-\delta$, for a fixed $\delta \in (0, 1/2)$ and all large n. W_n is the Wilks likelihood ratio statistics described as (8.135). Then

$$\frac{\log W_n - \mu_n}{\sigma_n} \xrightarrow{d} \mathcal{N}(0,1) \tag{8.136}$$

as $n \to \infty$, where

$$\mu_n = -c \log\left(1 - \frac{p}{n-1}\right) + \sum_{i=1}^{k} c_i \log\left(1 - \frac{p_i}{n-1}\right)$$

$$\sigma_n^2 = -\log\left(1 - \frac{p}{n-1}\right) + \sum_{i=1}^{k} c_i \log\left(1 - \frac{p_i}{n-1}\right)$$

with $c = p - n + \frac{3}{2}, c_i = p_i - n + \frac{3}{2}$.

In the theorem above, integers $k, p_1, \ldots, p_k$ and p all can depend on sample size n. The assumption that $\frac{\max_i p_i}{p} \leqslant 1-\delta$ rules out the situation where $\frac{\max_i p_i}{p} \to 1$ along the entire sequence or any subsequence.

8.9.8 Test of Mutual Dependence

A prominent feature of data collection nowadays is that the number of variables is comparable with the sample size. This is the opposite of the classical situations where many observations are made on low-dimensional data. Measuring mutual dependence is important in time-series analysis and cross-sectional panel-data analysis. While serial dependence can be characterized by the general spectral density function, mutual dependence is difficult to be described by a single criteria. This paper proposes a new statistic, due to [507], to test **mutual dependence** for a large number of high dimensional random vectors, including multiple time series and cross-sectional panel data.

Suppose that $\{X_{ji}\}, j = 1, \ldots, n, i = 1, \ldots, p$ are complex-valued random variables. For $1 \leq i \leq p$, let $\mathbf{x}_i = (X_{1i}, \ldots, X_{ni})^T$ denote the i-th time series and $\mathbf{x}_1, \ldots, \mathbf{x}_p$ be a panel of p time series, where n usually denotes the sample size in each of the time series data. In both theory and practice, it is common to assume that each of the time series $(X_{1i}, \ldots, X_{ni})$ is statistically *independent*, but it may be unrealistic to assume that $\mathbf{x}_1, \ldots, \mathbf{x}_p$ are independent or even uncorrelated. This is because there is no natural ordering for cross-sectional indices.

It may be necessary to test whether $\mathbf{x}_1, \ldots, \mathbf{x}_p$ are independent before a statistical model is used to model such data. The main motivation of including this section is to show how to use an empirical spectral distribution function-based test statistic for cross-sectional independence of $\mathbf{x}_1, \ldots, \mathbf{x}_p$. The aim is to test

$$\begin{aligned} &\mathcal{H}_0 : \mathbf{x}_1, \ldots, \mathbf{x}_p \text{ are independent} \\ &\mathcal{H}_1 : \mathbf{x}_1, \ldots, \mathbf{x}_p \text{ are not independent} \end{aligned} \tag{8.137}$$

where $\mathbf{x}_i = (X_{1i}, \ldots, X_{ni})^T$, for $i = 1, \ldots, p$.

Our approach essentially uses the characteristic function of the empirical spectral distribution of sample covariance matrices in large random-matrix theory. When $\mathbf{x}_1, \ldots, \mathbf{x}_p$ are mutual independent, the limiting spectral distribution (LSD) of the corresponding sample covariance matrix is the Marcenko–Pastur (M–P) law (see Section 3.5). From this point, any deviation of the LSD from M-P law is evidence of dependence. Indeed, Silverstein [175] and Bai and Zhou [508] report the LSD of the sample covariance matrix with correlations in columns and it is different from the M–P law. We need not draw observations again from the set of vectors of $\mathbf{x}_1, \ldots, \mathbf{x}_p$ due to the high dimensionality.

Assumption 8.9.14 For each $i = 1, \ldots, p$, $Y_{1i}, \ldots, Y_{ni}$ are independent and identically distributed (i.i.d) random variables with mean zero, variance one and finite fourth moment. When Y_{ji} are complex random variables we require $\mathbb{E}X_{ji}^2 = 0$. Let

$$\mathbf{x}_i = \boldsymbol{\Sigma}_n^{1/2} \mathbf{y}_i$$

with $\mathbf{y}_i = (Y_{1i}, \ldots, Y_{ni})^T$ and $\boldsymbol{\Sigma}_n^{1/2}$ being a Hermitian square root of the non-negative definite Hermitian matrix $\boldsymbol{\Sigma}_n$.

Assumption 8.9.15 $p = p(n)$ with $p/n \to c \in (0, \infty)$.

We stack p time series $\mathbf{x}_i$ one by one to form a data matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p) \in \mathbb{C}^{n \times p}$. Moreover, denote the sample covariance matrix by

$$\mathbf{S}_n = \frac{1}{n} \mathbf{X}^H \mathbf{X} \in \mathbb{C}^{p \times p}$$

where H stands for Hermitian transpose of the matrix $\mathbf{X}$. The empirical spectral distribution (ESD) of the sample covariance matrix $\mathbf{S}_n$ is defined as

$$F_{\mathbf{S}_n}(x) = \frac{1}{p} \sum_{i=1}^{p} \mathbb{I}\left(\lambda_i \leqslant x\right) \tag{8.138}$$

where $\lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_p$ are eigenvalues of $\mathbf{S}_n$.

It is well-known that if $\mathbf{x}_1, \ldots, \mathbf{x}_p$ are independent and $c_n = p/n \to c \in (0, \infty)$ then $F_{\mathbf{S}_n}(x)$ converges with probability one to the Marchenko–Pastur law $F_c(x)$ whose density has an explicit expression

$$f_c(x) = \begin{cases} \frac{1}{2\pi c}\frac{1}{x}\sqrt{(b-x)(a-x)}, & a \leqslant x \leqslant b \\ 0, & \text{otherwise} \end{cases} \tag{8.139}$$

and a point mass $1 - 1/c$ at the origin if $c > 1$, where $a = \left(1 - \sqrt{c}\right)^2$ and $b = \left(1 + \sqrt{c}\right)^2$.

When there is some correlation among $\mathbf{x}_1, \ldots, \mathbf{x}_p$, denoted by $\mathbf{\Sigma}_n$, the covariance matrix of the first row, $\mathbf{y}_1^T$, of $\mathbf{X}$. Then, under Assumption 8.9.14, when $F_{\mathbf{\Sigma}_n}(x) \xrightarrow{D} H(x)$, $F_{\mathbf{S}_n}(x)$ converges with probability one to a nonrandom distribution function $F_{c,H}(x)$ whose Stieltjes transform satisfies

$$m(z) = \int \frac{1}{x(1 - c - czm(z)) - z} dH(x) \tag{8.140}$$

The construction of our test statistic relies on the following observation: the limit of the ESD of the sample covariance matrix $\mathbf{S}_n$ is the M–P law by (8.139) when $\mathbf{x}_1, \ldots, \mathbf{x}_p$ are independent and satisfy Assumption 8.9.14 (Hypothesis $\mathcal{H}_0$), while the limit of the ESD is determined from (8.140) when there is some correlation among $\mathbf{x}_1, \ldots, \mathbf{x}_p$ with the covariance matrix $\mathbf{\Sigma}_p$ different from the identity matrix $\mathbf{I}_p$: $\mathbf{\Sigma}_p \neq \mathbf{I}_p$ (Hypothesis $\mathcal{H}_1$).

Moreover, preliminary investigations indicate that when $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are only uncorrelated (without any further assumptions), the limit of the ESD of $\mathbf{S}_n$ is not the M–P law (see [281]). These therefore motivate us to employ the ESD of $\mathbf{S}_n$, $F_{\mathbf{S}_n}(x)$, as a test statistic. There is no central limit theorem for $F_{\mathbf{S}_n}(x) - F_{c,H}(x)$ however, as argued in [188]. We instead consider the characteristic function of $F_{\mathbf{S}_n}(x)$.

The characteristic function of $F_{\mathbf{S}_n}(x)$ is

$$s_n(t) \triangleq \int e^{jtx} dF_{\mathbf{S}_n}(x) = \frac{1}{p}\sum_{i=1}^{p} e^{jt\lambda_i} \tag{8.141}$$

where $\lambda_i, i = 1, \ldots, p$ are eigenvalues of the sample covariance matrix of $\mathbf{S}_n$. Our test statistic is then proposed as follows:

$$M_n = \int_{T_1}^{T_2} \left| s_n(t) - s(t) \right| dU(t) \tag{8.142}$$

where $s(t) := s\left(t, c_n\right)$ is the characteristic function of $F_{c_n}(x)$, obtained from the M–P law $F_c(x)$ with c replaced by $c_n = p/n$, and $U(t)$ is a distributional function with its support on a compact interval, say $[T1, T2]$.

Assumption 8.9.16 Let $\mathbf{\Sigma}_p$ be a $p \times p$ random Hermitian non-negative definite matrix with a bounded spectral norm. Let $\mathbf{y}_j^T = \mathbf{z}_j^T \mathbf{\Sigma}_p^{1/2}$, where $\mathbf{\Sigma}_p$ satisfies $\left(\mathbf{\Sigma}_p^{1/2}\right)^2 = \mathbf{\Sigma}_p$ and $\mathbf{z}_j = \left(Z_{j1}, \ldots, Z_{jp}\right)^T, j = 1, \ldots, n$ are i.i.d random vectors, in which $Z_{ji}, j \leqslant n, i \leqslant p$ are i.i.d with mean zero, variance one and finite fourth moment.

The empirical spectral distribution $F_{\mathbf{\Sigma}_n}(x)$ of $\mathbf{\Sigma}_p$ converges weakly to a distribution H on $[0, \infty)$ as $n \to \infty$; all the diagonal elements of the matrix $\mathbf{\Sigma}_p$ are equal to 1.

Under Assumption 8.9.16, $\mathbf{S}_n$ becomes

$$\mathbf{S}_n = \boldsymbol{\Sigma}_p^{1/2} \mathbf{Z}^H \mathbf{Z} \boldsymbol{\Sigma}_p^{1/2} \tag{8.143}$$

where $\mathbf{Z} = \left(\mathbf{z}_1, \ldots, \mathbf{z}_n\right)^T$. Under Assumption 3, when $\boldsymbol{\Sigma}_p = \mathbf{I}_p$, the random vectors $\mathbf{x}_1, \ldots, \mathbf{x}_p$ are independent and when $\boldsymbol{\Sigma}_p \neq \mathbf{I}_p$, they are not independent.

To develop the asymptotic distribution of the test statistic, we introduce

$$G_n(x) = p\left[F_{\boldsymbol{\Sigma}_n}(x) - F_{c_n}(x)\right] \tag{8.144}$$

Then, $p\left(s_n(t) - s(t)\right)$ can be decomposed as a sum of the random part and the nonrandom part as follows:

$$\begin{aligned} p\left(s_n(t) - s(t)\right) &= \int e^{jtx} dG_n(x) \\ &= \int e^{jtx} d\left(p\left[F_{\boldsymbol{\Sigma}_n}(x) - F_{c_n,H_n}(x)\right]\right) \\ &\quad + \int e^{jtx} d\left(p\left[F_{c_n,H_n}(x) - F_{c_n}(x)\right]\right) \end{aligned} \tag{8.145}$$

where $F_{c_n,H_n}(x)$ is obtained from $F_{c,H}$ with c and H replaced by $c_n = p/n$ and $H_n = F_{\boldsymbol{\Sigma}_p}$.

Example 8.9.17 (a general panel data model) Consider a panel data model of the form

$$v_{ij} = w_{ij} + \frac{1}{\sqrt{p}} u_i, \quad i = 1, \ldots, p; j = 1, \ldots, n \tag{8.146}$$

where $\left\{w_{ij}\right\}$ is a sequence of i.i.d. real random variables with $\mathbb{E}w_{11} = 0$, and $\mathbb{E}w_{11}^2 = 1$, and u_i, $i = 1, \ldots, p$ are real random variables, and independent of $\left\{w_{ij}\right\}$, $i = 1, \ldots, p; j = 1, \ldots, n$.

For any $i = 1, \ldots, p$, set

$$\mathbf{v}_i = \left(v_{i1}, \ldots, v_{in}\right)^T \tag{8.147}$$

(8.146) can be written as

$$\mathbf{V} = \mathbf{W} + \mathbf{u}\mathbf{1}^T \tag{8.148}$$

where

$$\mathbf{V} = \left(\mathbf{v}_1, \ldots, \mathbf{v}_p\right)^T, \mathbf{u} = \left(\frac{1}{\sqrt{p}} u_1, \ldots, \frac{1}{\sqrt{p}} u_p\right)^T, \mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^p$$

Consider the sample covariance matrix

$$\mathbf{S}_n = \frac{1}{n} \mathbf{V}\mathbf{V}^T = \frac{1}{n} \left(\mathbf{W} + \mathbf{u}\mathbf{1}^T\right)\left(\mathbf{W} + \mathbf{u}\mathbf{1}^T\right)^T \tag{8.149}$$

By Lemma 5 of [507] and the fact that rank $\left(\mathbf{u}\mathbf{1}^T\right) \leqslant 1$, it can be concluded that the limit of the ESD of the matrix $\mathbf{S}_n$ is the same as that of the matrix $\frac{1}{n}\mathbf{W}\mathbf{W}^T$, i.e. the M-P law. Even so, it is desirable to use the proposed statistic M_n to test the null hypothesis of mutual independence.

For the model (8.146), in addition to Assumptions 8.9.14 and 8.9.15, we assume that

$$\mathbb{E}\,\|\mathbf{u}\|^4 < \infty, \quad \text{and} \quad \frac{1}{p^2}\mathbb{E}\left[\sum_{i\neq j}^{p}\left(u_i^2-\overline{u}\right)\left(u_j^2-\overline{u}\right)\right] \to 0 \quad \text{as } n\to\infty \tag{8.150}$$

where $\overline{u}$ is a positive constant. Then, the proposed test statistic p^2M_n converges in distribution to the random variable R_2 given by

$$R_2 = \int_{t_1}^{t_2}\left(|W(t)|^2 + |Q(t)|^2\right)dU(t) \tag{8.151}$$

where $(W(t), Q(t))$ is a Gaussian vector whose mean and covariance are specified in [507].

When $u_1,\ldots,u_p$ are independent and hence $\mathbf{v}_1,\ldots,\mathbf{v}_p$ are independent, condition (8.150) is true. □

8.9.9 Test of Presence of Spike Eigenvalues

We follow the notation of Section 8.9.8, unless defined otherwise.

Let $\mathbf{\Sigma}_p$ be a sequence of $p\times p$ nonrandom and non-negative definite Hermitian matrices. We consider the spiked population model introduced in [177] where the eigenvalues of the $\mathbf{\Sigma}_p$ are

$$\underbrace{a_1,\ldots,a_1}_{n_1},\ldots,\underbrace{a_k,\ldots,a_k}_{n_k},\underbrace{1,\ldots,1}_{p-M} \tag{8.152}$$

Here M and the multiplicity numbers (n_k) are fixed and satisfy $n_1+\cdots+n_k = M$. In other words, all the population eigenvalues are unit except some fixed number of them (the spikes). The model can be viewed as a finite-rank perturbation of the null case.

We analyze the effects caused by the spike eigenvalues on the fluctuations of linear spectral statistics of the form

$$T_n(f) = \sum_{i=1}^{p} f\left(\lambda_{n,i}\right) = F_{\mathbf{S}_n}(f) \tag{8.153}$$

where f is a given function. $\mathbf{S}_n$ is the sample covariance matrix defined in (8.143). As with the convergence of the spectral distributions, the presence of the spikes does not prevent a central limit theorem for $T_n(f)$; the centering term in the central limit theorem will, however, be modified according to the values of the spikes.

The spectral density H_n of $\mathbf{\Sigma}_n$ is

$$H_n = \frac{p-M}{p}\delta_1 + \frac{1}{p}\sum_{i=1}^{k} n_i\delta_{a_i} \tag{8.154}$$

The term

$$\frac{1}{p}\sum_{i=1}^{k} n_i\delta_{a_i}$$

vanishes when p tends to infinity, so it has no influence when considering limiting spectral distributions. However for the CLT, the term $pF_{c_n,H_n}(f)$ has a p in the front, and $\frac{1}{p}\sum_{i=1}^{k} n_i\delta_{a_i}$ is of order $O(1)$, and thus *cannot be neglected*.

In [160], a corrected likelihood ratio statistic L is proposed to test the hypothesis

$$\mathcal{H}_0 : \mathbf{\Sigma}_p = \mathbf{I}_p$$
$$\mathcal{H}_1 : \mathbf{\Sigma}_p \neq \mathbf{I}_p$$

They prove that, under $\mathcal{H}_0$, where

$$L = \mathrm{Tr}\left(\mathbf{S}_n\right) - \log\det\left(\mathbf{S}_n\right) - p$$
$$G_{c_n,H_n}\left(g\right) = 1 - \frac{c_n - 1}{c_n}\log\left(1 - c_n\right)$$
$$m\left(g\right) = -\frac{\log\left(1-c\right)}{2}$$
$$v\left(g\right) = -2\log\left(1-c\right) - 2c$$

At a significance level α (usually 0.05), the test will reject $\mathcal{H}_0$ when

$$L - pG_{c_n,H_n}\left(g\right) > m\left(g\right) + \Phi^{-1}\left(1-\alpha\right)\sqrt{v\left(g\right)}$$

where Φ is the standard normal cumulative distribution function.

However, the power function of this test remains unknown because the distribution of L under the general alternative hypothesis $\mathcal{H}_1$ is ill-defined. Consider the general case:

$$\begin{aligned}&\mathcal{H}_0 : \mathbf{\Sigma}_p = \mathbf{I}_p\\ &\mathcal{H}_1 : \mathbf{\Sigma}_p \text{ has the spiked structure (8.152)}\end{aligned} \tag{8.155}$$

In other words, we want to test the absence against the presence of possible spike eigenvalues in the population covariance matrix.

Recall that $F_{c_n,H_n}(x)$is defined below (8.145) from $F_{c,H}$ with c and H replaced by $c_n = p/n$ and $H_n = F_{\mathbf{\Sigma}_p}$.

Under the alternative $\mathcal{H}_1$ and for $f(x) = x - \log x - 1$ used in the statistic L, the centering term $F_{c_n,H_n}\left(f\right)$ can be found to be

$$1 + \frac{1}{p}\sum_{i=1}^{k} n_i a_i - \frac{M}{p} - \frac{1}{p}\sum_{i=1}^{k} n_i \log a_i - \left(1 - \frac{1}{c_n}\right)\log\left(1-c_i\right) + O\left(\frac{1}{n^2}\right)$$

due to the following formulas

$$F_{c_n,H_n}(x) = 1 + \frac{1}{p}\sum_{i=1}^{k} n_i \log a_i - \frac{M}{p} + O\left(\frac{1}{n^2}\right) \tag{8.156}$$

and

$$F_{c_n,H_n}\left(\log x\right) = \frac{1}{p}\sum_{i=1}^{k} n_i \log a_i - 1 + \left(1 - \frac{1}{c_n}\right)\log\left(1-c_i\right) + O\left(\frac{1}{n^2}\right) \tag{8.157}$$

Therefore we have found that under $\mathcal{H}_1$,

$$L - pF_{c_n,H_n}\left(f\right) \Rightarrow \mathcal{N}\left(m\left(g\right), v\left(g\right)\right)$$

It follows that the asymptotic power function of the test is

$$\beta(\alpha) = 1 - \Phi\left(\Phi^{-1}\left(1-\alpha\right) - \frac{\sum_{i=1}^{k} n_i\left(a_i - 1 - \log a_i\right)}{\sqrt{-2\log\left(1-c\right) - 2c}}\right) \tag{8.158}$$

Example 8.9.18 (spectrum sensing in cognitive radio network) Consider an application to spectrum sensing in a cognitive radio network, see, for example, [39]. Consider the problem

$$\mathcal{H}_0 : \mathbf{y} = \mathbf{w}$$
$$\mathcal{H}_1 : \mathbf{y} = \mathbf{s} + \mathbf{w}$$

where $\mathbf{w}$ is the white Gaussian random vector and $\mathbf{s}$, independent of $\mathbf{w}$, is the signal vector in presence. Then the true covariance matrix has the form

$$\mathcal{H}_0 : \sigma^2\mathbf{I}$$
$$\mathcal{H}_1 : \mathbf{\Sigma}_s + \sigma^2\mathbf{I}$$

where σ^2 is the noise variance and $\mathbf{\Sigma}_s$ is the covariance matrix of the signal vector. Using the eigenvalue decomposition

$$\mathbf{\Sigma}_x = \mathbf{U}\,\mathrm{diag}\left(\lambda_1, \lambda_2, \ldots, \lambda_p\right)\mathbf{U}^H, \qquad \mathbf{I} = \mathbf{U}\mathbf{U}^H$$

we have

$$\mathcal{H}_0 : \sigma^2\mathbf{I}$$
$$\mathcal{H}_1 : \mathbf{U}\,\mathrm{diag}\left(\lambda_1 + \sigma^2, \lambda_2 + \sigma^2, \ldots, \lambda_p + \sigma^2\right)\mathbf{U}^H$$

which is equivalent to (8.155). □

8.9.10 Large Dimension and Small Sample Size

The last few decades have seen explosive growth in data analysis, due to the rapid development of modern information technology. We are now in a setting where many very important data analysis problems are high dimensional. In many scientific areas the data dimension p can even be a lot larger than the sample size n. The main purpose of this section is to establish central limit theorems (CLTs) of linear functionals of eigenvlaues of sample covariance matrix when the dimension p is much larger than the sample size n, i.e., $p/n \to \infty$.

Consider the sample covariance matrix $\mathbf{S}_n = \frac{1}{n}\mathbf{X}_n\mathbf{X}_n^T$, where $\mathbf{X}_n = \left(X_{ij}\right)_{p\times n}$ and $X_{ij}, i = 1, \ldots, p, j = 1, \ldots, n$ are i.i.d. real random variables with mean zero and variance one. As we know, linear functionals of eigenvalues $\mathbf{S}_n$ are closely related to its empirical spectral distribution (ESD) function $F_{\mathbf{S}_n}(x)$. Here for any $n \times n$ Hermitian matrix $\mathbf{M}$ with real eigenvalues $\lambda_1, \ldots, \lambda_n$, its empirical spectral distribution of $\mathbf{M}$ is defined by

$$F_{\mathbf{M}}(x) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}\left(\lambda_j \leqslant x\right)$$

where $\mathbb{I}$ is the indicator function for the event $\left(\lambda_j \leqslant x\right)$. However, it is inappropriate to use $F_{\mathbf{S}_n}(x)$ when $p/n \to \infty$ because $\mathbf{S}_n$ has $(p-n)$ zero eigenvalues and hence $F_{\mathbf{S}_n}(x)$ converges to a degenerate distribution with probability one. Note that the eigenvalues of $\mathbf{S}_n$ are the same as those of $\frac{1}{n}\mathbf{X}_n^T\mathbf{X}_n$ except $(p-n)$ zero eigenvalues. Thus, instead, we turn to investigate the spectral of $\frac{1}{p}\mathbf{X}_n^T\mathbf{X}_n$ and renormalize it as

$$\mathbf{A} = \frac{1}{\sqrt{np}}\left(\mathbf{X}^T\mathbf{X} - p\mathbf{I}_n\right) \tag{8.159}$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix. When $p/n \to \infty$, under fourth moment conditions the central limit theorem (CLT) for linear spectral statistics (LSS) of $\mathbf{A}$ defined by the eigenvalues is established below.

The first breakthrough regarding $\mathbf{A}$ in (8.159) was made in Bai and Yin [509]. They proved with probability one

$$F_{\mathbf{A}}(x) \to F(x)$$

which is the so-called semicircle law with the density

$$F'(x) = \begin{cases} \frac{1}{2\pi}\sqrt{4-x^2}, & \text{if } |x| \leqslant 2 \\ 0 & \text{if } |x| > 2 \end{cases} \tag{8.160}$$

In random matrix theory, $F(x)$ is named as the limiting spectral distribution (LSD) of empirical spectral distribution (ESD) $F_{\mathbf{A}}(x)$.

In order to study the central limit theorem (CLT) of the linear functions of eigenvalues of $\mathbf{A}$, let $\mathcal{P}$ denote any open region on the real plane including $[2, 2]$, which is the support of $F(x)$, and $\mathcal{F}$ be the set of functions analytic on $\mathcal{P}$. For any $f \in \mathcal{F}$, define

$$Q_n(f) \triangleq n\int_{-\infty}^{+\infty} f(x)d\left(F_{\mathbf{A}}(x) - F(x)\right) - \frac{1}{\pi}\sqrt{\frac{n^3}{p}}\int_{-1}^{1} f(2x)\frac{4x^3-3x}{\sqrt{1-x^2}}dx \tag{8.161}$$

and its random part

$$Q_n^{(1)}(f) \triangleq n\int_{-\infty}^{+\infty} f(x)d\left(F_{\mathbf{A}}(x) - \mathbb{E}F_{\mathbf{A}}(x)\right) \tag{8.162}$$

Let$\{T_k\}$ be the family of Chebyshev polynomials, which is defined as

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x)$$

To give an alternative way of calculating the asymptotic covariance of $X(f)$ in Theorem 8.9.19 below, for any $f \in \mathcal{F}$ and any integer $k > 0$, we define

$$\begin{aligned}\Psi_k(f) &\triangleq \frac{1}{2\pi}\int_{-\pi}^{+\pi} f(2\cos\theta)e^{ik\theta}d\theta \\ &= \frac{1}{2\pi}\int_{-\pi}^{+\pi} f(2\cos\theta)\cos k\theta d\theta = \frac{1}{\pi}\int_{-1}^{+1} f(2x)T_k(x)\,dx\end{aligned}$$

The main result is formulated in the following theorem.

Theorem 8.9.19 Suppose that

(a) $\mathbf{X} = \left(X_{ij}\right)_{p\times n}$ where $\left\{X_{ij} : i = 1, 2, \ldots, p; j = 1, 2, \ldots, n\right\}$ are i.i.d. real random variables with $\mathbb{E}X_{11} = 0$, $\mathbb{E}X_{11}^2 = 1$, and $\nu_4 = \mathbb{E}X_{11}^4 < \infty$.

(b1) $n^3/p = O(1)$ as $n \to \infty$.

Then, for any $f_1, \ldots, f_k \in \mathcal{F}$, the finite dimensional random vector $(Q_n(f_1), \ldots, Q_n(f_k))$ converges weakly to a Gaussian vector $(X(f_1), \ldots, X(f_k))$ with mean function

$$\mathbb{E}X(f) = \frac{1}{\pi}\int_{-1}^{+1} f(2x)\left[2(\nu_4-3)x^3 - \left(\nu_4 - \frac{5}{2}\right)\right]\frac{1}{\sqrt{1-x^2}}dx + \frac{1}{4}(f(2)+f(-2)) \tag{8.163}$$

and variance function

$$\operatorname{cov}\left(X\left(f_1\right), X\left(f_2\right)\right) = \left(\nu_4 - 3\right)\Psi_1\left(f_1\right)\Psi_1\left(f_2\right) + 2\sum_{k=1}^{\infty} k\Psi_k\left(f_1\right)\Psi_k\left(f_2\right)$$
$$= \frac{1}{4\pi^2}\int_{-2}^{2}\int_{-2}^{2} f_1^{'}(x) f_2^{'}(y) H(x,y)\, dxdy \tag{8.164}$$

where

$$H(x,y) = \left(\nu_4 - 3\right)\sqrt{4-x^2}\sqrt{4-y^2} + 2\log\left(\frac{4 - xy + \sqrt{\left(4-x^2\right)\left(4-y^2\right)}}{4 - xy - \sqrt{\left(4-x^2\right)\left(4-y^2\right)}}\right)$$

If we interchange the roles of p and n, Birke and Dette [490] established the CLT for $Q_n(f)$ when $f(x) = x^2$ and $X_{ij} \sim \mathcal{N}(0,1)$.

Note that Theorem 8.9.19 is established under the restriction $n^3/p = O(1)$. The next theorem extends it to the general framework $n/p \to 0$. For this purpose, instead of using $Q_n(f)$ we define $G_n(f)$ as

$$G_n\left(f\right) \triangleq n\int_{-\infty}^{\infty} f(x)\, d\left(F_{\mathbf{\Lambda}}(x) - F(x)\right)$$
$$- \frac{n}{2\pi i}\oint_{|m|=\rho} f\left(-m - m^{-1}\right)\mathcal{X}_n(m)\frac{1-m^2}{m^2}dm \tag{8.165}$$

where

$$\mathcal{X}_n(m) \triangleq \frac{-\mathcal{B} + \sqrt{\mathcal{B}^2 - 4\mathcal{A}\mathcal{C}}}{2\mathcal{A}}, \quad \mathcal{A} = m - \sqrt{\frac{n}{p}}\left(1 + m^2\right)$$
$$\mathcal{B} = m^2 - 1 - \sqrt{\frac{n}{p}}m\left(1 + 2m^2\right), \quad \mathcal{C} = \frac{m^3}{n}\left(\frac{m^2}{1-m^2} + \nu_4 - 2\right) - \sqrt{\frac{n}{p}}m^4 \tag{8.166}$$

and $\sqrt{\mathcal{B}^2 - 4\mathcal{A}\mathcal{C}}$ is the complex number whose imaginary part has the same sign as the imaginary part of $\mathcal{B}$. The integral's contour is taken as $|m| = \rho$ with $\rho < 1$.

Theorem 8.9.20 Suppose that

(a) $\mathbf{X} = \left(X_{ij}\right)_{p\times n}$ where $\left\{X_{ij} : i = 1, 2, \ldots, p; j = 1, 2, \ldots, n\right\}$ are i.i.d. real random variables with $\mathbb{E}X_{11} = 0$, $\mathbb{E}X_{11}^2 = 1$ and $\nu_4 = \mathbb{E}X_{11}^4 < \infty$

(b2) $n/p \to 0$ as $n \to \infty$

Then, for any $f_1, \ldots, f_k \in \mathcal{F}$, the finite dimensional random vector $(G_n(f_1), \ldots, G_n(f_k))$ converges weakly to a Gaussian vector $(Y(f_1), \ldots, Y(f_k))$ with mean function $\mathbb{E}Y(f) = 0$ and covariance function $\operatorname{cov}(Y(f), Y(g))$ the same as that given in (8.164).

Therefore Theorem 8.9.19 is a special case of Theorem 8.9.20 when $n^3/p = O(1)$. The mean correction term in Theorem 8.9.19 has a simple and explicit expression and it vanishes when $f(x)$ is even or $n^3/p \to 0$. The proofs for the two theorems are almost the same.

If $n^3/p = O(1)$, Theorem 8.9.20 is consistent with Theorem 8.9.19. Indeed, since $n^3/p = O(1)$, we have $4\mathcal{AC} = o(1)$, $\mathcal{B} = m^2 - 1$. By (8.166)

$$n\mathcal{X}_n(m) = n \cdot \frac{-\mathcal{B} + \sqrt{\mathcal{B}^2 - 4\mathcal{AC}}}{2\mathcal{A}} = \frac{-2n\mathcal{C}}{\mathcal{B} + \sqrt{\mathcal{B}^2 - 4\mathcal{AC}}}$$
$$= \frac{m^2}{1-m^2}\left(\frac{m^2}{1-m^2} - \nu_4 - 2\right) + \sqrt{\frac{n^3}{p}}\frac{m^4}{1-m^2} + o(1)$$

So, by the same calculation as that in Section 5.1 in [510], we have

$$-\frac{n}{2\pi i}\oint_{|m|=\rho} f\left(-m - m^{-1}\right)\mathcal{X}_n(m)\frac{1-m^2}{m^2}dm$$
$$= -\frac{1}{2\pi i}\oint_{|m|=\rho} f\left(-m - m^{-1}\right)m\left[\frac{m^2}{1-m^2} - \nu_4 - 2 + \sqrt{\frac{n^3}{p}}\right]dm + o(1)$$
$$= -\left[\frac{1}{4}\left(f(2) + f(-2)\right) - \frac{1}{2}\Psi_0(f) + \left(\nu_4 - 3\right)\Psi_2(f)\right]$$
$$- \sqrt{\frac{n^3}{p}}\Psi_3(f) + o(1) \tag{8.167}$$

Example 8.9.21 (hypothesis test) Suppose that $\mathbf{y} = \mathbf{Hs}$ is a p-dimensional vector with covariance matrix $\mathbf{\Sigma} = \mathbf{HH}^T$ with $\mathbf{H}$ being a $p \times p$ matrix whose eigenvalues are positive and the entries of $\mathbf{s}$ being i.i.d random variables with mean zero and variance one. We want to test

$$\begin{aligned} &\mathcal{H}_0 : \mathbf{\Sigma}_p = \mathbf{I}_p \\ &\mathcal{H}_1 : \mathbf{\Sigma}_p \neq \mathbf{I}_p \end{aligned} \tag{8.168}$$

We are interested in (8.168) in the setting $p/n \to \infty$, for large p and small n. We often study the functions of the sample covariance matrix $\mathbf{S}_n$, especially the functions of its eigenvalues. Here we take $f(x) = x^2$ in (8.161) or (8.165). By Theorem 8.9.19 or Theorem 8.9.20, we then propose the test statistic as follows:

$$L_n = \frac{1}{2}\left[n\left(\int x^2 dF_{\mathbf{B}}(x) - \int x^2 dF(x)\right) - \left(\nu_4 - 2\right)\right]$$
$$= \frac{1}{2}\left(\operatorname{Tr}\mathbf{BB}^T - n - \left(\nu_4 - 2\right)\right) \tag{8.169}$$

where $\mathbf{B} = \sqrt{\frac{p}{n}}\left(\frac{1}{p}\mathbf{X}^T\mathbf{X} - \mathbf{I}_n\right)$ and $\mathbf{Y} = \left(\mathbf{y}_1, \ldots, \mathbf{y}_n\right)$. Since $\mathbf{H}^T\mathbf{H} = \mathbf{I_p}$ is equivalent to $\mathbf{HH}^T = \mathbf{I_p}$, under the null hypothesis $\mathcal{H}_0$, we have

$$L_n \xrightarrow{d} \mathcal{N}(0,1) \tag{8.170}$$

□

For comparison, let us introduce the work of [511] on the asymptotic power of likelihood ratio test (LRT) for the identity test in the setting when the dimension p is large compared to the sample size n. A natural approach to test (8.168) is to conduct estimations for some distance measures between $\mathbf{\Sigma}_p$ and $\mathbf{I}_p$ and there are two types of measures

that are widely used in literature. The first is based on the likelihood function, also called Stein's loss function:

$$L_{\text{Stein}}\left(\mathbf{\Sigma}_p\right) = \text{Tr}\left(\mathbf{\Sigma}_p\right) - \log\det\left(\mathbf{\Sigma}_p\right) - p \tag{8.171}$$

and the second is based on quadratic loss function:

$$L_{\text{Quad}}\left(\mathbf{\Sigma}_p\right) = \text{Tr}\left(\mathbf{\Sigma}_p - \mathbf{I}_p\right)^2 \tag{8.172}$$

To relax the Gaussian assumptions, we assume that the observations $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ satisfy a multivariate model

$$\mathbf{x}_i = \mathbf{\Sigma}_p^{1/2}\mathbf{z}_i + \boldsymbol{\mu}, \quad i = 1, \ldots, n \tag{8.173}$$

where $\boldsymbol{\mu}$ is a p-dimensional constant vector and the entries of $\mathbf{Z}_n = \left(Z_{ij}\right)_{p\times n} = \left(\mathbf{z}_1, \ldots, \mathbf{z}_n\right)$ are i.i.d. with

$$\mathbb{E}Z_{ij} = 0, \quad \mathbb{E}Z_{ij}^2 = 1, \quad \text{and} \quad \mathbb{E}Z_{ij}^4 = 4 + \Delta$$

The sample covariance matrix $\mathbf{S}_n$ is defined using

$$\mathbf{S}_n = \frac{1}{n-1}\sum_{k=1}^{n}\left(\mathbf{x}_k - \overline{\mathbf{x}}\right)\left(\mathbf{x}_k - \overline{\mathbf{x}}\right)^T$$

where $\overline{\mathbf{x}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$

Writing $c_n = p/n < 1$, the likelihood ratio test (LRT) statistic is defined as

$$L_n = \frac{1}{p}\text{Tr}\left(\mathbf{S}_n\right) - \log\det\left(\mathbf{S}_n\right) - 1 - d\left(c_n\right) \tag{8.174}$$

where $d(x) = 1 + (1/x - 1)\log(1-x), 0 < x < 1$. Under the null hypothesis, Wang *et al.* [511] derived the following asymptotic normality of L_n by using random matrix theories.

Theorem 8.9.22 When $\mathbf{\Sigma}_p = \mathbf{I}_p$, and $c_n = p/n \to c \in (0,1)$

$$\frac{pL_n - \mu_n}{\sigma_n} \xrightarrow{D} \mathcal{N}(0,1)$$

where

$$\mu_n = c_n(\Delta/2 - 1) - 3/2\log\left(1 - c_n\right), \quad \sigma_n^2 = -2c_n - 2\log\left(1 - c_n\right)$$

and $\xrightarrow{D}$ denotes convergence in distribution.

When $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ be i.i.d. multivariate normal distributions $\mathcal{N}\left(\boldsymbol{\mu}, \mathbf{\Sigma}_p\right)$ where $\Delta = 0$, Jiang *et al.* [473] derived a similar result as Theorem 8.9.22 by using the Selberg integral and they also considered the special situation where $p/n \to 1$. Based on the asymptotic normality under the respective null hypothesis, an asymptotic level α test based on L_n is given by

$$\phi = \mathbb{I}\left(\frac{pL_n - \mu_n}{\sigma_n} > z_{1-\alpha}\right) \tag{8.175}$$

where $z_{1-\alpha}$ denotes the $100 \times (1-\alpha)$-th percentile of the standard normal distribution. In the following theorem, we establish the convergence of L_n under the alternative $\mathcal{H}_1 : \mathbf{\Sigma}_p \neq \mathbf{I}_p$.

Theorem 8.9.23 When $\mathrm{Tr}\left(\mathbf{\Sigma}_p - \mathbf{I}_p\right)^2/p \to 0$ and $c_n = p/n \to c \in (0,1)$, we have

$$\frac{pL_n - L_{\mathrm{Stein}}\left(\mathbf{\Sigma}_p\right) - \mu_n}{\sigma_n} \xrightarrow{D} \mathcal{N}(0,1)$$

where $\mu_n = c_n(\Delta/2 - 1) - 3/2\log\left(1-c_n\right), \quad \sigma_n^2 = -2c_n - 2\log\left(1-c_n\right)$

In particular, when $L_{\mathrm{Stein}}\left(\mathbf{\Sigma}_p\right)$ tends to a constant, we have the following result.

Theorem 8.9.24 When $K_2\left(\mathbf{\Sigma}_p\right) \to b \in (0,\infty)$ and $c_n = p/n \to c \in (0,1)$, we have

$$\lim_{n\to\infty} \mathbb{P}_{\mathbf{\Sigma}_p}\left(\phi \text{ rejects } \mathcal{H}_0\right) = 1 - \Phi\left(z_{1-\alpha} - \frac{b}{\sqrt{-2c - 2\log(1-c)}}\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

It can be seen from Theorems 8.9.23 and 8.9.24 that the expression

$$1 - \Phi\left(z_{1-\alpha} - \frac{L_{\mathrm{Stein}}\left(\mathbf{\Sigma}_p\right)}{\sigma_n}\right) \tag{8.176}$$

gives good approximation to the power of the test in (8.175) until the power is extremely close to 1. In particular, when $L_{\mathrm{Stein}}\left(\mathbf{\Sigma}_p\right)$ is large, the power of the test ϕ will be close to 1 and it is hard for ϕ to distinguish between the two hypotheses if $L_{\mathrm{Stein}}\left(\mathbf{\Sigma}_p\right)$ tends to zero.

To derive the asymptotic power, a special covariance matrix was used in [484] and [512] as follows

$$\mathbf{\Sigma}_p^\star = \mathbf{I}_p + h\sqrt{\frac{p}{n}}\mathbf{v}\mathbf{v}^T \tag{8.177}$$

where h is a constant and $\mathbf{v}$ is an arbitrarily fixed unit vector. Here, by Theorem 8.9.24, we know that the asymptotic power of the LRT test (8.174) is

$$1 - \Phi\left(z_{1-\alpha} - \frac{h\sqrt{c} - \log\left(1 + h\sqrt{c}\right)}{\sqrt{-2c_n - 2\log\left(1-c_n\right)}}\right)$$

Therefore, compared with the tests based on $L_{\mathrm{Quad}}\left(\mathbf{\Sigma}_p\right)$ ([469, 476, 484]) whose power for $\mathbf{\Sigma}_p^\star$ is $1 - \Phi\left(z_{1-\alpha} - h^2/2\right)$, LRT is more sensitive to small eigenvalues ($h < 0$), not any bigger than one ($h > 0$). In particular, when $1 + h\sqrt{c}$ is close to 0, that is $\mathbf{\Sigma}_p^\star$ has a very small eigenvalue, the power will tend to 1.

8.10 Roy's Largest Root Test

In this section relatively accurate expressions for the distribution of Roy's largest root test were derived in the extreme setting of a rank-one concentrated noncentrality matrix. Deriving such expressions, even in this restricted case, has been an open problem in multivariate analysis for several decades and has potentially limited the practical use of Roy's test. The new distributions derived in [513] are simple and straightforward to compute. Moreover, as shown in the simulations, for small sample sizes and strong signals, they provide much more accurate expressions for the distribution of the largest root, compared to the classical Gaussian approximation. This section is also motivated for massive MIMO in Section 15.3.

First we consider an example for the motivation.

Example 8.10.1 (multiple response linear regression) Consider a linear model with n observations on an $m-$variate response

$$\mathbf{Y} = \mathbf{XB} + \mathbf{Z} \tag{8.178}$$

where $\mathbf{Y}$ is $n \times m$ and the known design matrix $\mathbf{X}$ is $n \times p$, so that the unknown coefficient matrix $\mathbf{B}$ is $p \times m$. Assume that $\mathbf{X}$ has full rank p. The Gaussian noise $\mathbf{Z}$ is assumed to have independent rows, each with mean zero and covariance $\mathbf{\Sigma}$, thus $\mathbf{Z} \sim \mathcal{N}\left(0, \mathbf{I}_n \otimes \mathbf{\Sigma}\right)$, where $\otimes$ represents the Kronecker product.

A common null hypothesis is $\mathbf{CB} = 0$. This is used, for example, to test (differences among) subsets of coefficients. We assume that the "contrast" matrix $\mathbf{C}$ has full rank $r \leq p$. Generalizing the univariate F test, it is traditional to form "hypothesis" and "error" sums of squares and cross products matrices, which under our Gaussian assumptions have independent Wishart distributions:

$$\mathbf{H} = \mathbf{Y}^T \mathbf{P}_H \mathbf{Y} \sim W_m\left(n_H, \mathbf{\Sigma}, \mathbf{\Omega}\right)$$
$$\mathbf{E} = \mathbf{Y}^T \mathbf{P}_E \mathbf{Y} \sim W_m\left(n_H, \mathbf{\Sigma}\right)$$

$\mathbf{P}_E$ is orthogonal projection of rank $n_E = n - p$ onto the error subspace, $\mathbf{P}_H$ is orthogonal projection of rank $n_E = r$ onto the hypothesis subspace for $\mathbf{CB}$, and $\mathbf{\Omega}$ is the noncentrality matrix corresponding to the regression mean $\mathbb{E}\mathbf{Y} = \mathbf{XB}$.

Classical tests use the eigenvalues of the F-like matrix $\mathbf{E}^{-1}\mathbf{H}$; our interest here is with Roy's largest root test, which is based on the largest of the eigenvalues, $\ell_1\left(\mathbf{E}^{-1}\mathbf{H}\right)$. Our approximation, valid for the case of rank one noncentrality matrix, employs the linear combination of two independent F distributions, one of which is noncentral.

Proposition 8.10.2 Suppose that $\mathbf{H} \sim W_m\left(n_H, \mathbf{\Sigma}, \mathbf{\Omega}\right)$, and $\mathbf{E} \sim W_m\left(n_H, \mathbf{\Sigma}\right)$ are independent Wishart matrices with $m > 1$ and $\nu = n_E - m > 1$. Assume that the noncentrality matrix has rank one, $\mathbf{\Omega} = \omega \mathbf{\Sigma}^{-1} \mathbf{v}\mathbf{v}^T$, for $\omega > 0$ and $\mathbf{v}$ of length one. If m, n_H, and n_E remains fixed and $\omega \to \infty$, then

$$\ell_1\left(\mathbf{E}^{-1}\mathbf{H}\right) \approx c_1 F_{a_1, b_1}(\omega) + c_2 F_{a_2, b_2}(\omega) + c_3 \tag{8.179}$$

where the F-variates are independent, and the numerator and denominator degrees of freedom are given by

$$a_1 = n_H, \quad b_1 = \nu + 1, \quad a_2 = m - 1, \quad b_2 = \nu + 2 \tag{8.180}$$
$$c_1 = a_1/b_1, \quad c_2 = a_2/b_2, \quad c_3 = a_2/(\nu(\nu - 1)) \tag{8.181}$$

∎

Consider a measurement system consisting of m sensors (antennas, smart meters, PMUs, etc). A standard model for the observed samples in the presence of a single signal is

$$\mathbf{x} = \sqrt{\rho_s}\xi\mathbf{h} + \sigma\mathbf{n} \tag{8.182}$$

where $\mathbf{h}$ is an unknown m-dimensional vector, which is assumed to be fixed during the measurement time window, ξ is a random variable distributed $\mathcal{N}(0,1)$, ρ_s is the signal strength, σ is the noise level and $\mathbf{n}$ is a random noise vector that follows a multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$.

In this section, for the sake of simplicity, we assume real valued signals and noise. The complex-valued case can be handled in a similar manner. Let $\mathbf{x}_i \in \mathbb{R}^m$, for $i = 1, \ldots, n_H$, denote n_H i.i.d. observations from (8.182), and let $\frac{1}{n_H}\mathbf{H}$ denote their sample covariance matrix:

$$\mathbf{H} = \sum_{i=1}^{n_H} \mathbf{x}_i \mathbf{x}_i^T \sim W_m\left(n_H, \mathbf{\Sigma} + \mathbf{\Omega}\right) \tag{8.183}$$

where $\mathbf{\Omega} = \rho_s \mathbf{h}\mathbf{h}^T$ has rank one. A fundamental problem in statistical signal processing is to test $\mathcal{H}_0 : \rho_s = 0$, no signal present, versus $\mathcal{H}_1 : \rho_s > 0$. If the noise covariance matrix $\mathbf{\Sigma}$ is known, the observed data can be whitened by the transformation $\mathbf{\Sigma}^{-1/2}\mathbf{x}_i$. Standard detection schemes then depend on the eigenvalues of $\mathbf{\Sigma}^{-1}\mathbf{H}$.

The second important case assumes that the noise covariance matrix $\mathbf{\Sigma}$ is *arbitrary and unknown*, but we have additional "noise-only" observations $\mathbf{z}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ for $j = 1, \ldots, n_E$. It is then traditional to estimate the noise covariance by $\frac{1}{n_E}\mathbf{E}$, where

$$\mathbf{E} = \sum_{i=1}^{n_E} \mathbf{z}_i \mathbf{z}_i^T \sim W_m\left(n_E, \mathbf{\Sigma}\right) \tag{8.184}$$

and devise detection schemes using the eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$.

Let ℓ_1 be the largest eigenvalue of either $\mathbf{\Sigma}^{-1}\mathbf{H}$ or $\mathbf{E}^{-1}\mathbf{H}$, depending on the specific setting. Roy's test accepts the alternative if $\ell_1 > t(\alpha)$ where $t(\alpha)$ is the threshold corresponding to a false alarm or type I error rate of α. The probability of detection, or power of Roy's test is defined as

$$P_D = \mathbb{P}\left[\ell_1 > t(\alpha) \,|\, \mathcal{H}_1\right]$$

If the matrix $\mathbf{\Sigma}$ is assumed to be known, without loss of generality we assume that $\mathbf{\Sigma} = \mathbf{I}$ and study the largest eigenvalue of $\mathbf{H}$, instead of $\mathbf{E}^{-1}\mathbf{H}$.

Proposition 8.10.3 Let $\mathbf{H} \sim W_m\left(n_H, \sigma^2\mathbf{I} + \lambda_H \mathbf{v}\mathbf{v}^T\right)$ with $||\mathbf{v}|| = 1$ and let λ_{max} be its largest eigenvalue. Then, with (m, λ_H, n_H) fixed, as $\sigma \to 0$

$$\lambda_{\max} = \left(\lambda_H + \sigma^2\right)\chi^2_{n_H} + \chi^2_{m-1}\sigma^2 + \frac{\chi^2_{m-1}\chi^2_{n_H-1}}{\left(\lambda_H + \sigma^2\right)\chi^2_{n_H}}\sigma^4 + o_p\left(\sigma^4\right) \tag{8.185}$$

where the three chi-square variates $\chi^2_{n_H}$, χ^2_{m-1} and $\chi^2_{n_H-1}$ are independent.

Approximations to the moments of $\lambda_{\max}$ follow directly. From (8.185), independence of the chi-square variates and $\mathbb{E}\left(1/\chi_n^2\right) = 1/(n-2)$, we obtain

$$\mathbb{E}\lambda_{\max} \approx n_H\lambda_H + \left(m-1+n_H\right)\sigma^2 + \frac{(m-1)\left(n_H-1\right)}{\left(\lambda_H+\sigma^2\right)\left(n_H-2\right)}\sigma^4 \tag{8.186}$$

It is natural to set $\omega = \lambda_H n_H$. Set $\sigma = 1$ and suppose that $\lambda_H = \omega/n_H$ is large. Then the variance of λ_{max}

$$\mathrm{Var}\left(\lambda_{\max}\right) = 2n_H\lambda_H^2 + 4n_H\lambda_H + 2\left(m-1+n_H\right) + o(1) \tag{8.187}$$

In the joint limit $m \to \infty, n_H \to \infty$ with $m/n_H \to c > 0$, there is a large recent literature in random matrix theory on the behavior of the "spiked model," beginning for example with [335]. The basic phenomenon is a phase transition at $\lambda = \sqrt{c}$ (for $\sigma = 1$): for $\lambda < \sqrt{c}$, $\ell_1(\mathbf{H})$ has asymptotically a Tracy–Widom distribution with zero power, while $\lambda > \sqrt{c}$, $\ell_1(\mathbf{H})$ follows an approximate Gaussian distribution with different scaling and asymptotic power one. We will see that in the fixed (m, n_H) cases we consider, corresponding to $\lambda > \sqrt{c}$, the Gaussian approximation is typically inferior to the ones developed here.

Next, we consider the two matrix case, where $\mathbf{\Sigma}$ is unknown and estimated from data. The following proposition considers the signal detection setting under the alternative hypothesis of a single Gaussian signal present.

Proposition 8.10.4 Suppose that $\mathbf{H} \sim W_m\left(n_H, \sigma^2\mathbf{I} + \lambda_H\mathbf{v}\mathbf{v}^T\right)$ and $\mathbf{E} \sim W_m\left(n_E, \mathbf{I}\right)$ are independent Wishart matrices, with $m > 1$ and $||\mathbf{v}|| = 1$. If m, n_H and n_E remain fixed and $\lambda_H \to \infty$, then

$$\ell_1\left(\mathbf{E}^{-1}\mathbf{H}\right) \approx c_1\left(1+\lambda_H\right)F_{a_1,b_1} + c_2F_{a_2,b_2} + c_3 \tag{8.188}$$

where the F-variates are independent, and with $\nu = n_E - m > 1$, the numerator and denominator degrees of freedom are given by (8.180) and (8.181).

When the covariance is unknown, we have

$$\mathbb{E}\ell_1\left(\mathbf{E}^{-1}\mathbf{H}\right) \approx \frac{1}{n_E-m-1}\left[\left(\lambda_H+1\right)n_H + m - 1\right] \tag{8.189}$$

Let $\hat{\mathbf{\Sigma}} = \frac{1}{n_E}\mathbf{E}$ be an unbiased estimator of $\mathbf{\Sigma}$. Comparison with Proposition 8.10.3 shows that $\mathbb{E}\ell_1\left(\hat{\mathbf{\Sigma}}^{-1}\mathbf{H}\right)$ exceeds $\mathbb{E}\ell_1\left(\mathbf{\Sigma}^{-1}\mathbf{H}\right)$ by a multiplicative factor $\frac{n_E}{n_E-m-1}$, so that the largest eigenvalue of $n_E\mathbf{E}^{-1}\mathbf{H}$ is thus typically larger than that of the matrix $\mathbf{\Sigma}^{-1}\mathbf{H}$.

Nadakuditi and Silverstein [514] studied the limiting value (but not the distribution) of the largest eigenvalue of $\left(n_E/n_H\right)\ell_1\left(\mathbf{E}^{-1}\mathbf{H}\right)$ in the limit $m, n_E, n_H \to \infty$ with $m/n_E \to c_E, m/n_H \to c_E$ (also in non-Gaussian cases). It can be verified that, in this limit, our formula (8.189) agrees with the large λ_H limit of their expression to leading order terms. Hence, our analysis shows that their limiting expressions (Eq. (23)) are in fact quite accurate for the mean of $\ell_1\left(\mathbf{E}^{-1}\mathbf{H}\right)$, even at relatively small values of m, n_E, n_H.

8.11 Optimal Tests of Hypotheses for Large Random Matrices

Theorem 8.11.1 (Neyman–Pearson Theorem [515]) Let $X_1, X_2, \ldots, X_n$, where n is a fixed positive integer, denote a random sample from a distribution that has pdf or pmf $f(x;\theta)$. Then the likelihood of $X_1, X_2, \ldots, X_n$ is

$$L(\theta;\mathbf{x}) = \prod_{i=1}^{n} f\left(x_i;\theta\right), \quad \text{for } \mathbf{x}' = \left(x_1, \ldots, x_n\right)$$

Let θ' and θ'' be distinct values of θ so that $\Omega = \{\theta : \theta = \theta', \theta''\}$, and let k be a positive number. Let C be a subset of the sample space such that:

(a) $\frac{L(\theta';\mathbf{x})}{L(\theta'';\mathbf{x})} \leqslant k$, for each point $\mathbf{x} \in C$
(b) $\frac{L(\theta';\mathbf{x})}{L(\theta'';\mathbf{x})} \geqslant k$, for each point $\mathbf{x} \in C$
(c) $\alpha = P_{\mathcal{H}_0}[\mathbf{X} \in C]$

Then C is a best critical region of size α for testing the simple hypothesis $\mathcal{H}_0 : \theta = \theta'$ against the alternative simple hypothesis $\mathcal{H}_1 : \theta = \theta''$.

Example 8.11.2 (likelihood ratio test formulated with random vectors [515]) Let $\mathbf{X}' = \left(X_1, \ldots, X_n\right)$ denote a random sample from the distribution that has the pdf

$$f(x;\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2}\right), \quad -\infty < x < \infty$$

We want to test the simple hypothesis $\mathcal{H}_0 : \theta = \theta' = 0$ against the alternative simple hypothesis $\mathcal{H}_1 : \theta = \theta'' = 1$. Now the likelihood ratio rest is

$$\frac{L(\theta';\mathbf{x})}{L(\theta'';\mathbf{x})} = \frac{\prod_{i=1}^{n} f\left(x_i;\theta'\right)}{\prod_{i=1}^{n} f\left(x_i;\theta''\right)} = \frac{\left(1/\sqrt{2\pi}\right)^n \exp\left(-\sum_{i=1}^{n} x^2/2\right)}{\left(1/\sqrt{2\pi}\right)^n \exp\left(-\sum_{i=1}^{n} (x-1)^2/2\right)}$$
$$= \exp\left(-\sum_{i=1}^{n} x_i + \frac{n}{2}\right) \qquad \square$$

We assume that the observations space corresponds to a set of n observations: $\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \ldots, \mathbf{r}_n$. Thus each set can be thought of as a point in the n-dimensional space and can be denoted by a vector

$$\mathbf{r} \triangleq \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \vdots \\ \mathbf{r}_n \end{bmatrix}$$

In the binary hypothesis problem we know that either $\mathcal{H}_0$ or $\mathcal{H}_1$ is true. The likelihood ratio is denoted by $\Lambda(\mathbf{R})$

$$\Lambda(\mathbf{R}) \triangleq \frac{p_{\mathbf{r}||\mathcal{H}_1}\left(\mathbf{R}|\mathcal{H}_1\right)}{p_{\mathbf{r}|\mathcal{H}_0}\left(\mathbf{R}|\mathcal{H}_0\right)} \tag{8.190}$$

where **R** is a random vector while **r** represents one realization of the random vector **R**. The Bayes criterion leads us to a likelihood ratio test

$$\Lambda(\mathbf{R}) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \eta \tag{8.191}$$

Because the natural logarithm is a monotonic function, and both sides of (8.191) are positive, an equivalent test is the log-likelihood ratio test

$$\ln \Lambda(\mathbf{R}) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \ln \eta \tag{8.192}$$

Example 8.11.3 (likelihood ratio test formulated with scalar valued random variables [516]) We consider an example. We assume that under $\mathcal{H}_1$ the source output is a constant voltage A. Under $\mathcal{H}_0$, the source output is zero. Before observation, the voltage is corrupted by an additive noise. We sample the output waveform each second and obtain N samples. Each noise sample n_i is a zero-mean Gaussian random variable n with variance σ^2. The noise samples at various instants are independent random variables and are independent of the source output. The observations under the two hypotheses are

$$\begin{aligned} &\mathcal{H}_1 : r_i = A + n_i \qquad i = 1, 2, \ldots, N \\ &\mathcal{H}_0 : r_i = n_i \qquad\quad i = 1, 2, \ldots, N \end{aligned} \tag{8.193}$$

and

$$p_{n_i}(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \tag{8.194}$$

because the noise samples are Gaussian.

The probability density of r_i under each hypothesis follows readily:

$$p_{\mathbf{r}|\mathcal{H}_1}(\mathbf{R}|\mathcal{H}_1) = p_{n_i}(R_i - A) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(R_i - A)^2}{2\sigma^2}\right) \tag{8.195}$$

and

$$p_{\mathbf{r}|\mathcal{H}_0}(\mathbf{R}|\mathcal{H}_0) = p_{n_i}(R_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{R_i^2}{2\sigma^2}\right) \tag{8.196}$$

Because the n_i are statistically independent, the joint probability density of the r_i (or, equivalently, of the vector **r**) is simply the product of the individual probability densities. Thus

$$p_{\mathbf{r}|\mathcal{H}_1}(\mathbf{R}|\mathcal{H}_1) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(R_i - m)^2}{2\sigma^2}\right) \tag{8.197}$$

and

$$p_{\mathbf{r}|\mathcal{H}_0}(\mathbf{R}|\mathcal{H}_0) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{R_i^2}{2\sigma^2}\right) \tag{8.198}$$

Inserting into (8.190), we have

$$\Lambda(\mathbf{R}) = \frac{p_{\mathbf{r}||\mathcal{H}_1}(\mathbf{R}|\mathcal{H}_1)}{p_{\mathbf{r}|\mathcal{H}_0}(\mathbf{R}|\mathcal{H}_0)} = \frac{\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(R_i - A)^2}{2\sigma^2}\right)}{\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{R_i^2}{2\sigma^2}\right)} \tag{8.199}$$

After canceling common terms and taking the logarithm, we have

$$\ln \Lambda(\mathbf{R}) = \frac{A}{\sigma^2} \sum_{i=1}^{N} R_i - \frac{NA^2}{2\sigma^2} \tag{8.200}$$

Thus the likelihood ratio test is

$$\frac{A}{\sigma^2} \sum_{i=1}^{N} R_i - \frac{NA^2}{2\sigma^2} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \ln \eta$$

or, equivalently

$$\sum_{i=1}^{N} R_i \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \frac{\sigma^2}{A} \ln \eta + \frac{NA}{2} \triangleq \gamma \tag{8.201}$$

We see that the processor simply adds the observations and compares them with a threshold. □

Example 8.11.4 (likelihood ratio test formulated with vector valued random variables [516]) A set of scalar-valued random variables $r_1, r_2, \ldots, r_N$ is defined as jointly Gaussian if all their linear combinations are Gaussian random variables.

A vector $\mathbf{r}$ is a Gaussian random vector when its components $r_1, r_2, \ldots, r_N$ are jointly Gaussian random variables.

In words, if

$$z = \sum_{i=1}^{N} g_i r_i \triangleq \mathbf{G}^T \mathbf{r} \tag{8.202}$$

If we define

$$\mathbb{E}[\mathbf{r}] = \mathbf{m} \tag{8.203}$$

and

$$\mathrm{Cov}[\mathbf{r}] = \mathbb{E}\left[(\mathbf{r} - \mathbf{m})\left((\mathbf{r} - \mathbf{m})^T\right)\right] \triangleq \mathbf{\Sigma} \tag{8.204}$$

then (8.202) implies that the characteristic function (Fourier transform) of $\mathbf{r}$ is

$$M_{\mathbf{r}}(j\mathbf{v}) \triangleq \mathbb{E}\left[e^{j\mathbf{v}^T\mathbf{r}}\right] = \exp\left(+j\mathbf{v}^T\mathbf{m} - \tfrac{1}{2}\mathbf{v}^T\mathbf{\Sigma}\mathbf{v}\right) \tag{8.205}$$

and assuming $\mathbf{\Sigma}$ is nonsingular, the probability density of $\mathbf{r}$ is

$$p_{\mathbf{r}}(\mathbf{R}) = \left[(2\pi)^{N/2}(\det \mathbf{\Sigma})^{1/2}\right]^{-1} \exp\left[-\tfrac{1}{2}(\mathbf{R} - \mathbf{m})^T \mathbf{\Sigma}(\mathbf{R} - \mathbf{m})\right] \tag{8.206}$$

A hypothesis-testing problem is called a general Gaussian problem if $p_{\mathbf{r}|\mathcal{H}_0}(\mathbf{R}|\mathcal{H}_0)$ and $p_{\mathbf{r}|\mathcal{H}_1}(\mathbf{R}|\mathcal{H}_1)$ are Gaussian densities. An estimation problem is called a general Gaussian problem if $p_{\mathbf{r}|\mathbf{b}}(\mathbf{R}|\mathbf{B})$ for all $\mathbf{B}$.

Let us focus on the binary hypothesis testing of the general Gaussian problem. We assume that the observations space is N-dimensional. Points in the space are denoted by the N-point vector (or column matrix) $\mathbf{r}$:

$$\mathbf{r} = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_N \end{pmatrix}$$

Under the first hypothesis $\mathcal{H}_1$, we assume that $\mathbf{r}$ is a Gaussian random vector, which is completely specified by its mean vector and covariance matrix. We denote these quantities as

$$\mathbb{E}\left[\mathbf{r}|\mathcal{H}_1\right] = \begin{pmatrix} \mathbb{E}\left[r_1|\mathcal{H}_1\right] \\ \mathbb{E}\left[r_2|\mathcal{H}_1\right] \\ \vdots \\ \mathbb{E}\left[r_N|\mathcal{H}_1\right] \end{pmatrix} \triangleq \begin{pmatrix} m_{11} \\ m_{21} \\ \vdots \\ m_{N1} \end{pmatrix} \triangleq \mathbf{m}_1 \tag{8.207}$$

The covariance matrix is

$$\mathbf{K}_1 \triangleq \mathbb{E}\left[\left(\mathbf{r}-\mathbf{m}_1\right)\left(\left(\mathbf{r}-\mathbf{m}_1\right)^T\right)|\mathcal{H}_1\right] \tag{8.208}$$

The probability density of $\mathbf{r}$ on $\mathcal{H}_1$,

$$p_{\mathbf{r}|\mathcal{H}_1}\left(\mathbf{R}|\mathcal{H}_1\right) = \left[(2\pi)^{N/2}\left(\det\mathbf{K}_1\right)^{1/2}\right]^{-1}\exp\left[-\tfrac{1}{2}\left(\mathbf{R}-\mathbf{m}_1\right)^T\mathbf{K}_1^{-1}\left(\mathbf{R}-\mathbf{m}_1\right)\right] \tag{8.209}$$

Similarly, we have for $\mathcal{H}_0$

$$p_{\mathbf{r}|\mathcal{H}_0}\left(\mathbf{R}|\mathcal{H}_0\right) = \left[(2\pi)^{N/2}\left(\det\mathbf{K}_0\right)^{1/2}\right]^{-1}\exp\left[-\tfrac{1}{2}\left(\mathbf{R}-\mathbf{m}_0\right)^T\mathbf{K}_0^{-1}\left(\mathbf{R}-\mathbf{m}_0\right)\right] \tag{8.210}$$

where

$$\mathbf{K}_0 \triangleq \mathbb{E}\left[\left(\mathbf{r}-\mathbf{m}_0\right)\left(\left(\mathbf{r}-\mathbf{m}_0\right)^T\right)|\mathcal{H}_0\right]$$

Using the definition of (8.190), the likelihood ratio test follows easily

$$\begin{aligned}\Lambda(\mathbf{R}) &\triangleq \frac{p_{\mathbf{r}|\mathcal{H}_1}\left(\mathbf{R}|\mathcal{H}_1\right)}{p_{\mathbf{r}|\mathcal{H}_0}\left(\mathbf{R}|\mathcal{H}_0\right)} \\ &= \frac{\left[(2\pi)^{N/2}\left(\det\mathbf{K}_1\right)^{1/2}\right]^{-1}\exp\left[-\tfrac{1}{2}\left(\mathbf{R}-\mathbf{m}_1\right)^T\mathbf{K}_1^{-1}\left(\mathbf{R}-\mathbf{m}_1\right)\right]}{\left[(2\pi)^{N/2}\left(\det\mathbf{K}_0\right)^{1/2}\right]^{-1}\exp\left[-\tfrac{1}{2}\left(\mathbf{R}-\mathbf{m}_0\right)^T\mathbf{K}_0^{-1}\left(\mathbf{R}-\mathbf{m}_0\right)\right]} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \eta\end{aligned} \tag{8.211}$$

Taking the logarithms, we obtain

$$\begin{aligned}&\tfrac{1}{2}\left(\mathbf{R}-\mathbf{m}_1\right)^T\mathbf{K}_1^{-1}\left(\mathbf{R}-\mathbf{m}_1\right)-\tfrac{1}{2}\left(\mathbf{R}-\mathbf{m}_0\right)^T\mathbf{K}_0^{-1}\left(\mathbf{R}-\mathbf{m}_0\right)\\&\underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}}\ln\eta+\tfrac{1}{2}\ln\left(\det\mathbf{K}_1\right)-\tfrac{1}{2}\ln\left(\det\mathbf{K}_0\right)\triangleq\gamma\ *\end{aligned}\tag{8.212}$$

We see that the test consists of finding the difference between two *quadratic* forms.

Let us now consider the repeated measurements. The hypothesis testing problem is expressed as

$$\begin{aligned}&\mathcal{H}_0:\mathbf{y}_i=\mathbf{z}_i, &&\mathbf{y}_i\in\mathbb{C}^{p\times1},\quad \mathbf{m}_i\in\mathbb{C}^{p\times1},\quad \mathbf{z}_i\in\mathbb{C}^{p\times1}\\&\mathcal{H}_1:\mathbf{y}_i=\mathbf{m}_i+\mathbf{z}_i, &&i=1,2,\ldots,N\end{aligned}\tag{8.213}$$

for complex Gaussian noise vectors $\mathbf{z}_i\sim\mathcal{CN}\left(\mathbf{0},\sigma^2\mathbf{I}_p\right)$, $i=1,2,\ldots,N$, the pdf of a random vector is

$$p_{\mathbf{z}_i}(\mathbf{x})=\frac{1}{(\pi\sigma^2)^p}\exp\left[-\frac{1}{\sigma^2}\mathbf{x}^H\mathbf{x}\right]$$

The joint pdf of N random vectors is given by

$$p_{\mathbf{z}_1}\left(\mathbf{x}_1\right)p_{\mathbf{z}_2}\left(\mathbf{x}_2\right)\cdots p_{\mathbf{z}_N}\left(\mathbf{x}_N\right)=\frac{1}{(\pi\sigma^2)^{Np}}\prod_{i=1}^{N}\exp\left[-\frac{1}{\sigma^2}\mathbf{x}_i^H\mathbf{x}_i\right]$$

The likelihood ratio is

$$\Lambda=\frac{p_{\mathcal{H}_1}\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N\right)}{p_{\mathcal{H}_0}\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N\right)}=\frac{\prod_{i=1}^{N}\exp\left[-\frac{1}{\sigma^2}\left(\mathbf{y}_i-\mathbf{m}_i\right)^H\left(\mathbf{y}_i-\mathbf{m}_i\right)\right]}{\prod_{i=1}^{N}\exp\left[-\frac{1}{\sigma^2}\mathbf{y}_i^H\mathbf{y}_i\right]}$$

The likelihood ratio test is

$$\Lambda\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N\right)\underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}}\eta$$

The log-likelihood ratio test becomes

$$\ln\Lambda\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N\right)=-\frac{1}{\sigma^2}\sum_{i=1}^{N}\left(\mathbf{y}_i-\mathbf{m}_i\right)^H\left(\mathbf{y}_i-\mathbf{m}_i\right)+\frac{1}{\sigma^2}\sum_{i=1}^{N}\mathbf{y}_i^H\mathbf{y}_i\underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}}\ln\eta$$

By eliminating the non-data-dependent terms we have

$$2\,\mathrm{Re}\left(\sum_{i=1}^{N}\mathbf{m}_i^H\mathbf{y}_i\right)\underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}}\sigma^2\ln\eta+\sum_{i=1}^{N}\mathbf{m}_i^H\mathbf{m}_i\tag{8.214}$$

This is the replica correlator for complex vector data. □

Before we present the likelihood ratio test in terms of large random matrices, we need to recall some known results in random matrix theory. The whole matrix $\mathbf{X}$ is treated as an element in the matrix probability space.

Now we have prepared all the ingredients to present the core results in this section, which appear here for the first time, to the best of our knowledge. Our goal is to make a parallel development with scalar and vector random variables of Van Tree. Here, we deal with matrix-valued random variables in the matrix probability space.

Example 8.11.5 (likelihood ratio test formulated with Gaussian random matrices) The observation is corrupted by an additive noise. We sample the data each second and obtain N matrix-valued samples. The entries of each noise sample $\mathbf{Z}_i$, which is matrix-valued, are zero-mean Gaussian random variables with variance σ^2. The noise samples at various instants are independent (matrix-valued) random variables and are independent of the source input. The observations under each hypothesis are

$$\begin{aligned} \mathcal{H}_0 &: \mathbf{Y}_i = \quad \mathbf{Z}_i, \qquad\qquad \mathbf{Z}_i \in \mathbb{C}^{p\times n}, \qquad i = 1, \dots, N \\ \mathcal{H}_1 &: \mathbf{Y}_i = \mathbf{M}_i + \mathbf{Z}_i, \quad \mathbf{M}_i \in \mathbb{C}^{p\times n}, \quad \mathbf{Z}_i \in \mathbb{C}^{p\times n}, \qquad i = 1, \dots, N \end{aligned} \tag{8.215}$$

where $\mathbf{M}_i$ may be a random or deterministic matrix, and from (3.43), we have

$$p_{\mathbf{Z}_i}(\mathbf{X}) = c \exp\left(-\frac{1}{\sigma^2} \operatorname{Tr}\left(\mathbf{X}^H \mathbf{X}\right)\right), \quad i = 1, \dots, N \tag{8.216}$$

because the entries of all the noise samples are Gaussian.

The probability density of $\mathbf{Y}_i$ under each hypothesis follows readily:

$$\mathcal{H}_1 : \; p_{\mathbf{Z}_i}\left(\mathbf{Y}_i - \mathbf{M}_i\right) = c \exp\left(-\frac{1}{\sigma^2} \operatorname{Tr}\left(\left(\mathbf{Y}_i - \mathbf{M}_i\right)^H \left(\mathbf{Y}_i - \mathbf{M}_i\right)\right)\right) \tag{8.217}$$

and

$$\mathcal{H}_0 : \quad p_{\mathbf{Z}_i}\left(\mathbf{Y}_i\right) = c \exp\left(-\frac{1}{\sigma^2} \operatorname{Tr}\left(\mathbf{Y}_i^H \mathbf{Y}_i\right)\right) \tag{8.218}$$

Because the $\mathbf{Z}_i$ are statistically independent, the joint probability density of all the N random matrices $\mathbf{Z}_i, i = 1, 2, \dots, N$ is simply the product of the independent individual probability density functions. See the derivation of (3.36). See also Example 3.9.4, in particular (3.45) and (3.46). Thus

$$\mathcal{H}_1 : \; p_1\left(\mathbf{Y}_i - \mathbf{M}_i\right) = c^N \prod_{i=1}^{N} \exp\left(-\frac{1}{\sigma^2} \operatorname{Tr}\left(\left(\mathbf{Y}_i - \mathbf{M}_i\right)^H \left(\mathbf{Y}_i - \mathbf{M}_i\right)\right)\right) \tag{8.219}$$

and

$$\mathcal{H}_0 : \quad p_0\left(\mathbf{Y}_i\right) = c^N \prod_{i=1}^{N} \exp\left(-\frac{1}{\sigma^2} \operatorname{Tr}\left(\mathbf{Y}_i^H \mathbf{Y}_i\right)\right) \tag{8.220}$$

Using the likelihood ratio test principle, in analogy with (8.191), we have

$$\Lambda = \frac{\prod_{i=1}^{N} \exp\left(-\frac{1}{\sigma^2} \operatorname{Tr}\left(\left(\mathbf{Y}_i - \mathbf{M}_i\right)^H \left(\mathbf{Y}_i - \mathbf{M}_i\right)\right)\right)}{\prod_{i=1}^{N} \exp\left(-\frac{1}{\sigma^2} \operatorname{Tr}\left(\mathbf{Y}_i^H \mathbf{Y}_i\right)\right)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \eta \tag{8.221}$$

Taking the logarithm to both sides, we have

$$\ln \Lambda = -\frac{1}{\sigma^2} \sum_{i=1}^{N} \operatorname{Tr}\left(\left(\mathbf{Y}_i - \mathbf{M}_i\right)^H \left(\mathbf{Y}_i - \mathbf{M}_i\right)\right) + \frac{1}{\sigma^2} \sum_{i=1}^{N} \operatorname{Tr}\left(\mathbf{Y}_i^H \mathbf{Y}_i\right) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \ln \eta$$

Canceling the common term, and with some simplification, we obtain

$$\frac{1}{N}\sum_{i=1}^{N}\left[\mathrm{Tr}\left(\mathbf{Y}_i^H\mathbf{M}_i\right)+\mathrm{Tr}\left(\mathbf{M}_i{}^H\mathbf{Y}_i\right)\right]\underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}}\frac{\sigma^2}{N}\ln\eta+\sum_{i=1}^{N}\mathrm{Tr}(\mathbf{M}_i^H\mathbf{M}_i) \tag{8.222}$$

Since $\mathrm{Tr}\left(\mathbf{A}^H\right)=(\mathrm{Tr}\,\mathbf{A})^*$, where z^* is the conjugate of a complex number z, identifying $\mathbf{A}=\mathbf{Y}_i^H\mathbf{M}_i$ gives

$$\frac{1}{N}\sum_{i=1}^{N}\left[\langle\mathbf{Y}_i,\mathbf{M}_i\rangle+\langle\mathbf{Y}_i,\mathbf{M}_i\rangle^*\right]\underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}}\frac{\sigma^2}{N}\ln\eta+\langle\mathbf{M}_i,\mathbf{M}_i\rangle \tag{8.223}$$

where the notation $\langle\mathbf{A},\mathbf{B}\rangle=\mathrm{Tr}\left(\mathbf{A}^H\mathbf{B}\right)$ represents the inner product (scalar product) of two matrices. For square $n\times n$ Hermitian matrices $\mathbf{Y}_i^H=\mathbf{Y}$, and $\mathbf{M}_i^H=\mathbf{M}_i$, we have

$$\frac{1}{N}\sum_{i=1}^{N}\langle\mathbf{Y}_i,\mathbf{M}_i\rangle\underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}}\frac{\sigma^2}{2N}\ln\eta+\frac{1}{2}\langle\mathbf{M}_i,\mathbf{M}_i\rangle \tag{8.224}$$

For $n=1$, we have $\mathbf{M}_i=\mathbf{m}_i\in\mathbb{C}^{p\times1}$, $\mathbf{Y}_i=\mathbf{y}_i\in\mathbb{C}^{p\times1}$, where random matrices reduce to random vectors (or rank-1 matrices). For this case, from (8.223) we obtain

$$2\,\mathrm{Re}\sum_{i=1}^{N}\mathbf{m}_i\mathbf{y}_i\underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}}\sigma^2\ln\eta+\sum_{i=1}^{N}\mathbf{m}_i^H\mathbf{m}_i \tag{8.225}$$

which is exactly (8.214) for random vectors. The difference between (8.225) and (8.222) is fundamental. To be explicit, we rewrite (8.222) as the sum of random matrices

$$\mathrm{Tr}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{Y}_i^H\mathbf{M}_i\right)+\mathrm{Tr}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{M}_i{}^H\mathbf{Y}_i\right)\underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}}\frac{\sigma^2}{N}\ln\eta+\mathrm{Tr}\left(\sum_{i=1}^{N}\mathbf{M}_i^H\mathbf{M}_i\right)$$

or

$$2\,\mathrm{Re}\left\{\mathrm{Tr}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{Y}_i^H\mathbf{M}_i\right)\right\}\underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}}\frac{\sigma^2}{N}\ln\eta+\mathrm{Tr}\left(\sum_{i=1}^{N}\mathbf{M}_i^H\mathbf{M}_i\right)$$

Concentration of spectral measure, unique to high dimensions, can be exploited when large random matrices are summed up. See Qiu and Wicks [40] for a treatment of this topic. The trace function of the sum of random matrices is of basic role.

For a special case $\mathbf{M}_i=A,\mathbf{Y}_i=Y_i$, which are scalar random variables for both, we have

$$\sum_{i=1}^{N}Y_i\underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}}\frac{\sigma^2}{2A}\ln\eta+\frac{1}{2}NA \tag{8.226}$$

which is identical to (8.201), the case for scalar random variables. The minor difference between (8.201) and (8.226) stems from a difference in the convention of σ^2, (a factor of 2), in (8.194) and (8.216).

Using the linearity of the trace, (8.222) can be expressed as

$$\mathrm{Tr}\left(\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{Y}_i\right)^H\mathbf{M}_i\right)+\mathrm{Tr}\left(\mathbf{M}_i^H\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{Y}_i\right)\right)\underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}}\frac{\sigma^2}{N}\ln\eta+\mathrm{Tr}(\mathbf{M}_i^H\mathbf{M})$$

Writing

$$\overline{\mathbf{Y}}=\frac{1}{N}\sum_{i=1}^{N}\mathbf{Y}_i$$

we have

$$\mathrm{Tr}\left(\overline{\mathbf{Y}}^H\mathbf{M}_i\right)+\mathrm{Tr}\left(\mathbf{M}_i^H\overline{\mathbf{Y}}\right)\underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}}\frac{\sigma^2}{N}\ln\eta+\mathrm{Tr}(\mathbf{M}_i^H\mathbf{M}_i)$$

or

$$\left\langle\overline{\mathbf{Y}},\mathbf{M}_i\right\rangle+\left\langle\overline{\mathbf{Y}},\mathbf{M}_i\right\rangle^*\underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}}\frac{\sigma^2}{N}\ln\eta+\mathrm{Tr}(\mathbf{M}_i^H\mathbf{M}_i)$$

For $\mathcal{H}_1$, we are able to move $\mathbf{M}_i$ to the left side to obtain $\mathbf{Z}_i=\mathbf{Y}_i-\mathbf{M}_i$. We can use free probability theory to extend from the above $\mathbf{M}_i$ to a large random matrix $\mathbf{A}_i$; thus we have

$$\mathcal{H}_1:\mathbf{Y}_i=\mathbf{A}_i+\mathbf{Z}_i,\qquad\mathbf{A}_i\in\mathbb{C}^{p\times n},\quad\mathbf{Z}_i\in\mathbb{C}^{p\times n},\quad i=1,\ldots,N$$

Non-Hermitian random matrices in free probability (see Chapter 6) can be used. In this generalized case, we use the asymptotically free random variables

$$\mathbf{Z}_i=\mathbf{Y}_i\boxminus\mathbf{A}_i \tag{8.227}$$

where $\boxminus$ denotes the free deconvolution [517].

For more generalized density, from (3.44), we have

$$p_{\mathbf{Z}_i}(\mathbf{X})=c\exp\left(-\mathrm{Tr}\,V\left(\mathbf{X}^H\mathbf{X}\right)\right) \tag{8.228}$$

For illustration, we only consider the special case of Gaussian matrices in the above.

In practice, we muse estimate $\mathbf{M}$ in (8.215).

Now the observations under each hypothesis are

$$\begin{aligned}&\mathcal{H}_0:\quad\mathbf{Y}_i=\mathbf{Z}_i,\quad\mathbf{Z}_i\in\mathbb{C}^{p\times n},\quad i=1,\ldots,N\\&\mathcal{H}_1:\quad\mathbf{Y}_i=\sqrt{SNR}\mathbf{A}_i+\mathbf{Z}_i,\quad\mathbf{A}_i\in\mathbb{C}^{p\times n},\quad\mathbf{Z}_i\in\mathbb{C}^{p\times n},\quad i=1,\ldots,N\end{aligned} \tag{8.229}$$

where $\mathbf{A}_i$ is a random matrix, independent of $\mathbf{Z}_i$, and SNR represents the signal-to-noise ratio. Here $\mathbf{A}_i$ and $\mathbf{Z}_i$ are two independent Gaussian random matrices given by (8.216). The log-likelihood ratio test becomes

$$\begin{aligned}\ln\Lambda=&-\frac{1}{\sigma^2}\mathrm{Tr}\left(\sum_{i=1}^{N}\left(\mathbf{Y}_i-\sqrt{\mathrm{SNR}}\mathbf{A}_i\right)^H\left(\mathbf{Y}_i-\sqrt{\mathrm{SNR}}\mathbf{A}_i\right)\right)\\&+\frac{1}{\sigma^2}\mathrm{Tr}\left(\sum_{i=1}^{N}{\mathbf{Y}_i}^H\mathbf{Y}_i\right)\underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}}\ln\eta\end{aligned} \tag{8.230}$$

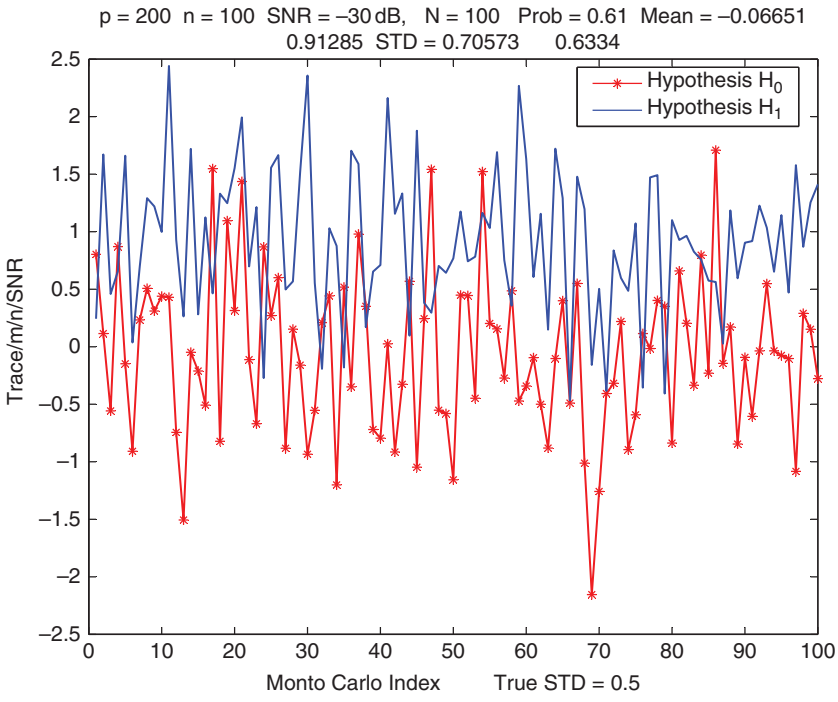


Figure 8.1 Log-likelihood functions defined in (8.230) under hypothesis $\mathcal{H}_0$ and hypothesis $\mathcal{H}_1$ for different Monte Carlo realizations. $p = 200$, $n = 100$, SNR $= -30$dB, N=100.

Hypothesis $\mathcal{H}_0$ can be viewed as the case SNR $= 0$. Our intuition leads us to study what is the minimum value of SNR, such that, on the left-hand-side of (8.230), the first term (hypothesis $\mathcal{H}_1$) can be found to be different from the second term (hypothesis $\mathcal{H}_0$). Figure 8.1 illustrates this intuition. In [40, p. 494], by empirical argument, we have obtained metric functions for hypothesis testing, similar to (8.230). The MATLAB code given below has been used to generate Figure 8.1.

```
clear all;
m=200;n=100; N=100;N_try=100;
SNR_dB=-30;
SNR=10^(SNR_dB/10)

Concentration=0;
STD= 0.38395;
t=1;

%**********************

for H1=0:1
```

```
Number=0;
for i_try=1:N_try                  %  Hypothesis Testing

MatrixAB=zeros(n,n);

for i=1:N                          %Monte Carlo for Expectation
X=1/sqrt(2)*randn(m,n)+sqrt(-1)/sqrt(2)*randn(m,n);
X1=1/sqrt(2)*randn(m,n)+sqrt(-1)/sqrt(2)*randn(m,n);
if H1==1
C=(X-sqrt(SNR)*X1)'*(X-sqrt(SNR)*X1);     % H1 SNR
else
C=X'*X;                                   % H0 SNR=0
end

MatrixAB=MatrixAB+C; % H1: Signal plus noise H0: noise only
end

Expectation_MatrixAB=MatrixAB/N; % expectation
Metric_Trace=real(trace((Expectation_MatrixAB)));
   % covariance
Metric_Trace=Metric_Trace-m*n;
Metric_Trace=(Metric_Trace)/m/n/SNR % mn

if abs(Metric_Trace- Concentration)> t*STD
 Number=Number+1 % H1
end
Record_Trace(i_try,H1+1)=Metric_Trace;
end %i_try
MEAN=sum(Record_Trace)/N_try;

Mean_Metric=MEAN
Prob=Number/N_try
STD=std(Record_Trace);
STD_True=1/m/n/SNR*sqrt(N);
end %H

p=m;
figure(1)
plot(1:N_try,Record_Trace(1:N_try,1),'r-*',1:N_try,
   Record_Trace(1:N_try,2),'b')
xlabel(['Monto Carlo Index' '
   True STD=' num2str(STD_True) ])
ylabel('Trace/m/n/SNR ')
legend('Hypothesis H_0','Hypothesis H_1')
title(['p=' num2str(p) ' n=' num2str(n) '
   SNR = ' num2str(10*log10(SNR)) 'dB,\ldots
N=' num2str(N) ' Prob=' num2str(Prob) '
```

```
   Mean=' num2str(MEAN) ' STD=' num2str(STD) ])
grid
```

□

Example 8.11.6 (likelihood ratio test formulated with Wishart random matrices) We refer to Example 3.9.2 for Wishart random matrices. For an $n \times m$ complex Gaussian matrix

$$p(\mathbf{X}) = \frac{1}{\pi^{nm}} \exp\left(-\operatorname{Tr} \mathbf{X}^H \mathbf{X}\right)$$

Following (3.39), the p.d.f. of $\mathbf{A} = \mathbf{X}^H \mathbf{X}$ as

$$p(\mathbf{A}) = \frac{1}{C_{\beta,n}} \exp\left(-\frac{\beta}{2} \operatorname{Tr} \mathbf{A}\right) (\det \mathbf{A})^{\beta/2(n-m+1-2/\beta)} \tag{8.231}$$

where $C_{\beta,n}$ is a normalization constant.

Consider the hypothesis-testing problem

$$\begin{aligned} &\mathcal{H}_0 : \mathbf{R} = \mathbf{A} \\ &\mathcal{H}_1 : \mathbf{R} = \mathbf{B} + \mathbf{A}, \quad \mathbf{A} \in \mathbb{C}^{m \times m}, \quad \mathbf{B} \in \mathbb{C}^{m \times m} \end{aligned} \tag{8.232}$$

where $\mathbf{B}$ is a deterministic matrix with $\mathbf{R} - \mathbf{B} > 0$. Obviously, we have $\mathbf{A} \geq 0$, because $\mathbf{A} = \mathbf{X}^H \mathbf{X}$. But in general, $\mathbf{A}$ is a large random matrix.

Using (8.231), the likelihood ratio is

$$\Lambda(\mathbf{A}) = \frac{p_1(\mathbf{A})}{p_0(\mathbf{A})} = \frac{p_1(\mathbf{R}-\mathbf{B})}{p_0(\mathbf{R})} = \frac{\exp\left(-\frac{\beta}{2} \operatorname{Tr}(\mathbf{R}-\mathbf{B})\right) (\det(\mathbf{R}-\mathbf{B}))^{\beta/2(n-m+1-2/\beta)}}{\exp\left(-\frac{\beta}{2} \operatorname{Tr} \mathbf{R}\right) (\det \mathbf{R})^{\beta/2(n-m+1-2/\beta)}}$$

The likelihood ratio test is given by

$$\Lambda(\mathbf{A}) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \eta$$

or equivalently

$$\ln \Lambda(\mathbf{A}) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \ln \eta$$

We have

$$\begin{aligned} \ln \Lambda(\mathbf{A}) &= \frac{\beta}{2}\left(n-m+1-\frac{2}{\beta}\right) \ln\det(\mathbf{R}-\mathbf{B}) - \frac{\beta}{2} \operatorname{Tr}(\mathbf{R}-\mathbf{B}) \\ &\quad - \frac{\beta}{2}\left(n-m+1-\frac{2}{\beta}\right) \ln\det(\mathbf{R}) + \frac{\beta}{2} \operatorname{Tr} \mathbf{R} \\ &= \frac{\beta}{2}\left(n-m+1-\frac{2}{\beta}\right) \ln \frac{\det(\mathbf{R}-\mathbf{B})}{\det(\mathbf{R})} + \frac{\beta}{2} \operatorname{Tr} \mathbf{B} \end{aligned}$$

For an $m \times m$ positive definitive matrix $\mathbf{C} > 0$, following (3.18), we have

$$\log\det(\mathbf{C}) = \operatorname{Tr} \log(\mathbf{C})$$

Using this relation, we obtain

$$\log \frac{\det(\mathbf{R}-\mathbf{B})}{\det(\mathbf{R})} = \operatorname{Tr}\log(\mathbf{R}-\mathbf{B}) - \operatorname{Tr}\log(\mathbf{R}) = \operatorname{Tr}\log\left(\mathbf{R}^{-1}(\mathbf{R}-\mathbf{B})\right)$$
$$= \operatorname{Tr}\log\left(\mathbf{I}-\mathbf{R}^{-1}\mathbf{B}\right)$$

where $\mathbf{I}-\mathbf{R}^{-1}\mathbf{B} > 0$ implied by the assumption $\mathbf{R}-\mathbf{B} > 0$. Finally, we have

$$\operatorname{Tr}\log\left(\mathbf{I}-\mathbf{R}^{-1}\mathbf{B}\right) + \frac{\beta}{2}\operatorname{Tr}(\mathbf{B}) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \frac{2}{\beta\left(n-m+1-\frac{2}{\beta}\right)}\ln\eta \tag{8.233}$$

where $\mathbf{I}-\mathbf{R}^{-1}\mathbf{B} > 0$, as pointed out before.

Let us inspect the test metric in (8.233). We have the form of

$$\operatorname{Tr} f(\mathbf{Y}) = \sum_{i=1}^{m} f\left(\lambda_i(\mathbf{Y})\right)$$

where $\mathbf{Y} > 0$ is a large random matrix and $f(x)$ is a convex function. In (8.233), $f(x) = \ln(1-x)$ for $x < 1$ is a convex function. The concentration of spectral measure phenomenon, unique to problems in high dimensions, is relevant here. See [40] for a systemic treatment.

See also Example 3.6.3 for the form of

$$\mathbb{E}\left[\operatorname{Tr}\left(f\left(\mathbf{X}\mathbf{Y}\mathbf{X}^H\right)\right)\right], \quad \mathbf{Y} > 0$$

where $\mathbf{X}$ is an $n \times n$ Gaussian random matrix with complex, independent, and identically distributed entries of zero mean and unit variance.

Now consider a sample problem:

$$\begin{aligned} &\mathcal{H}_0 : \mathbf{R}_i = \mathbf{A}_i \\ &\mathcal{H}_1 : \mathbf{R}_i = \mathbf{B}_i + \mathbf{A}_i, \qquad \mathbf{A} \in \mathbb{C}^{m\times m}, \qquad \mathbf{B}_i \in \mathbb{C}^{m\times m}, \quad i = 1,\ldots,N \end{aligned} \tag{8.234}$$

where $\mathbf{A}_i$ is given by (8.231) and $\mathbf{B}_i$s are random matrices with the assumption of $\mathbf{R}_i - \mathbf{B}_i > 0$. All the N random matrices $\mathbf{A}_i, i = 1,\ldots,N$ are independent from each other. All the N random matrices $\mathbf{B}_i, i = 1,\ldots,N$ are independent of each other. $\mathbf{A}_i, i = 1,\ldots,N$ are independent of $\mathbf{B}_i, i = 1,\ldots,N$. The joint pdf of all the N independent random matrices is the product of their individual pdfs:

$$p(\mathbf{A}_1)\,p(\mathbf{A}_2)\cdots p(\mathbf{A}_N) = \prod_{i=1}^{N} p(\mathbf{A}_i)$$
$$= \frac{1}{\left(C_{\beta,n}\right)^N}\prod_{i=1}^{N}\exp\left(-\frac{\beta}{2}\operatorname{Tr}\mathbf{A}_i\right)\left(\det\mathbf{A}_i\right)^{\beta/2(n-m+1-2/\beta)}$$
$$= \frac{1}{\left(C_{\beta,n}\right)^N}\exp\left(-\frac{\beta}{2}\sum_{i=1}^{N}\mathbf{A}_i\right)\left(\prod_{i=1}^{N}\det\mathbf{A}_i\right)^{\beta/2(n-m+1-2/\beta)}$$

Using (8.231), the likelihood ratio is

$$\begin{aligned} p\left(\mathbf{A}_1\right) p\left(\mathbf{A}_2\right) \cdots p\left(\mathbf{A}_N\right) &= \prod_{i=1}^{N} p\left(\mathbf{A}_i\right) \\ &= \frac{1}{\left(C_{\beta,n}\right)^N} \prod_{i=1}^{N} \exp\left(-\frac{\beta}{2} \operatorname{Tr} \mathbf{A}_i\right) \left(\det \mathbf{A}_i\right)^{\beta/2(n-m+1-2/\beta)} \\ &= \frac{1}{(C_{\beta,n})^N} \exp\left(-\frac{\beta}{2} \operatorname{Tr} \sum_{i=1}^{N} \mathbf{A}_i\right) \left(\prod_{i=1}^{N} \det \mathbf{A}_i\right)^{\beta/2(n-m+1-2/\beta)} \end{aligned}$$

The log-likelihood ratio is obtained as

$$\begin{aligned} \ln \Lambda &= -\frac{\beta}{2} \operatorname{Tr} \left[\sum_{i=1}^{N} \left(\mathbf{R}_i - \mathbf{B}_i\right) - \sum_{i=1}^{N} \mathbf{R}_i\right] \\ &\quad + \frac{\beta}{2}\left(n-m+1-\frac{2}{\beta}\right) \left[\sum_{i=1}^{N} \ln \det\left(\mathbf{R}_i - \mathbf{B}_i\right) - \sum_{i=1}^{N} \ln \det\left(\mathbf{R}_i\right)\right] \\ &= \frac{\beta}{2} \operatorname{Tr} \sum_{i=1}^{N} \mathbf{B}_i + \frac{\beta}{2}\left(n-m+1-\frac{2}{\beta}\right) \sum_{i=1}^{N} \ln \det\left(\mathbf{R}_i^{-1}\left(\mathbf{R}_i - \mathbf{B}_i\right)\right) \\ &= \frac{\beta}{2} \operatorname{Tr} \sum_{i=1}^{N} \mathbf{B}_i + \frac{\beta}{2}\left(n-m+1-\frac{2}{\beta}\right) \sum_{i=1}^{N} \ln \det\left(\mathbf{I} - \mathbf{R}_i^{-1}\mathbf{B}_i\right) \end{aligned}$$

where we have used the property $\det(\mathbf{AB}) = \det(\mathbf{A}) \det(\mathbf{B})$ in the second line. Sometimes the inverse of the random matrices $\mathbf{R}_i^{-1}$ do not exist. Using the relation $\log\det(\cdot) = \operatorname{Tr}\log(\cdot)$, we obtain

$$\ln \Lambda(\mathbf{A}) = \frac{\beta}{2} \operatorname{Tr} \sum_{i=1}^{N} \mathbf{B}_i + \frac{\beta}{2}\left(n-m+1-\frac{2}{\beta}\right) \sum_{i=1}^{N} \operatorname{Tr} \ln\left(\mathbf{I} - \mathbf{R}_i^{-1}\mathbf{B}_i\right)$$

which reduces to the case of $N = 1$ that we have studied before. Here $\mathbf{I} - \mathbf{R}_i^{-1}\mathbf{B}_i > 0$ implied by the assumption of $\mathbf{R}_i - \mathbf{B}_i > 0$. The sum of random matrices (appearing in the above equation) have been systemically studied in [40].

The log-likelihood ratio test becomes

$$\ln \Lambda(\mathbf{A}) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \ln \eta$$

□

The results in Example 8.11.5 and Example 8.11.6 appear novel, first obtained by the first author (Qiu) on February 21, 2014. He has been motivated to understand the finding in [61]: the trace function of random matrices performs the best among a number of algorithms. This deceptively simple finding has a far-reaching impact on his research program. A discussed in [40, Section 13.1], the trace function exploits the concentration of spectral measure phenomenon in a better manner than the matrix functions such the maximum or the minimum eigenvalue. The significance of the novel results is to justify

the empirical findings using the classical likelihood ratio test principle, by reformulating the problem in terms of large Gaussian random matrices.

We must treat the whole random matrix $\mathbf{X}$ as a whole element in some matrix probability space. This way we may use the probability density functions for a Gaussian random matrix (8.216) and a Wishart random matrix (8.231). Other generalized random matrices may be studied in this framework as well. As we pointed out previously, the power of random matrix theory is twofold. First, the eigenvalue distribution (empirical spectral measure) is universal: it is the same for many different distributions of the matrix entries. Second, we can treat the whole matrix as one element in some matrix probability space.

At an extremely low SNR of −30 dB, the noise is 1000 times larger than the signal, and any attempt to use eigenvectors and eigenvalues seems useless. Rather, our focus should be on the control of uncertainties (measured by variance) of the matrix functions that are used for hypothesis testing metrics. As a result, concentration of spectral measure—unique to high-dimensional problems—plays a central role in this framework. We are no longer satisfied with the assumption that a true covariance matrix can be estimated. Rather, we make a *direct* approach of formulating the problem in terms of large random matrices.

8.12 Matrix Elliptically Contoured Distributions

Matrix elliptically contoured distributions can be used to model data that are neither independent nor Gaussian.

The distribution of the sample covariance matrix, which has Wishart distribution [116], plays a central role in almost all multivariate inferential procedures. These techniques depend on functions of random matrices such as determinants, traces, and eigenvalues. Thus random matrices are the backbone of multivariate statistical analysis. Observed random phenomena often can be described by random matrices that include the dependence structure of the relevant random vectors.

Let $\mathbf{X}$ be a random matrix of dimension $p \times n$. Then, $\mathbf{X}$ is said to have a matrix variate elliptically contoured (m.e.c.) distribution if its characteristic functions have the form

$$\phi_{\mathbf{X}}(\mathbf{T}) = \text{etr}\left(j\mathbf{T}^T\mathbf{M}\right)\psi\left(\text{Tr}\left(\mathbf{T}^T\boldsymbol{\Sigma}\mathbf{T}\boldsymbol{\Phi}\right)\right)$$

with $\mathbf{T} : p \times n, \mathbf{M} : p \times n, \boldsymbol{\Sigma} : p \times p, \boldsymbol{\Phi} : n \times n, \boldsymbol{\Sigma} \geqslant 0, \boldsymbol{\Phi} \geqslant 0,$ and $\psi : [0, \infty) \to \mathbb{R}$. Here $\text{etr}(\cdot) = \exp(\text{Tr}(\cdot))$

This distribution will be denoted by $E_{p,n}(\mathbf{M}, \boldsymbol{\Sigma} \otimes \boldsymbol{\Phi}, \Psi)$.

For $n = 1$, we say that $\mathbf{X}$ has a vector variate elliptically contoured distribution. It is also called multivariate elliptical distribution. Then the characteristic function of $\mathbf{X}$ takes the form

$$\phi_{\mathbf{x}}(\mathbf{t}) = \exp\left(j\mathbf{t}^T\mathbf{m}\right)\psi\left(\mathbf{t}^T\mathbf{m}\right),$$

where $\mathbf{t}$ and $\mathbf{m}$ are p-dimensional vectors. In this case, in the notation $E_{p,n}(\mathbf{M}, \boldsymbol{\Sigma} \otimes \boldsymbol{\Phi}, \Psi)$, the index n can be dropped; that is, $E_p(\mathbf{m}, \boldsymbol{\Sigma}, \Psi)$ will denote the distribution $E_{p,1}(\mathbf{m}, \boldsymbol{\Sigma}, \Psi)$

Let $\mathbf{m}$ be a $p \times 1$ constant vector, and $\mathbf{A}$ be a $p \times p$ constant matrix. Random vector $\mathbf{x}$ is said to have an multivariate elliptic distribution with parameter $\mathbf{m}$ and $\boldsymbol{\Sigma} = \mathbf{A}^T\mathbf{A}$, if

it can be put in the form $\mathbf{x} = \mathbf{m} + \mathbf{Az}$, where $\mathbf{z}$ is a random vector following a spherical distribution.

The following three statements are equivalent.

- $E_{p,1}(\mathbf{m}, \mathbf{\Sigma}, \Psi)$.
- The probability density function of $\mathbf{x}$ is of the form $\frac{1}{\sqrt{\det \mathbf{\Sigma}}} g\left((\mathbf{x}-\mathbf{m})^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m})\right)$.
- The characteristic function of $\mathbf{x}$ is of the form $\exp\left(j\mathbf{t}^T\mathbf{m}\right)\Psi\left(\mathbf{t}^T\mathbf{\Sigma}\mathbf{t}\right)$.

The next result shows the relationship between matrix variate and vector variate elliptically contoured distributions. Let $\mathbf{X}$ be a $p \times n$ random matrix and $\mathbf{x} = vec\left(\mathbf{X}^T\right)$. Then, $\mathbf{X} \sim E_{p,n}(\mathbf{M}, \mathbf{\Sigma} \otimes \mathbf{\Phi}, \Psi)$ if and only if $\mathbf{x} \sim E_{p,n}\left(vec\left(\mathbf{M}^T\right), \mathbf{\Sigma} \otimes \mathbf{\Phi}, \Psi\right)$.

Linear functions of a random matrix with m.e.c. distribution also have elliptic contoured distributions. Let $\mathbf{X} \sim E_{p,n}(\mathbf{M}, \mathbf{\Sigma} \otimes \mathbf{\Phi}, \Psi)$. Assume $\mathbf{C} : q \times m, \mathbf{A} : q \times p$, and $\mathbf{B} : n \times m$ are constant matrices. Then,

$$\mathbf{AXB} + \mathbf{C} \sim E_{q,m}\left(\mathbf{AMB} + \mathbf{C}, \left(\mathbf{A}^T\mathbf{\Sigma}\mathbf{A}\right) \otimes \left(\mathbf{B}^T\mathbf{\Phi}\mathbf{B}\right), \Psi\right)$$

Proof: The characteristic function of $\mathbf{Y} = \mathbf{AXB} + \mathbf{C}$ can be written as

$$\begin{aligned}
\phi_{\mathbf{Y}}(\mathbf{T}) &\triangleq \mathbb{E}\left(\text{etr}\left(j\mathbf{T}^T\mathbf{Y}\right)\right) \\
&= \mathbb{E}\left(\text{etr}\left(j\mathbf{T}^T(\mathbf{AXB} + \mathbf{C})\right)\right) \\
&= \mathbb{E}\left(\text{etr}\left(j\mathbf{T}^T\mathbf{AXB}\right)\right)\text{etr}\left(j\mathbf{T}^T\mathbf{C}\right) \\
&= \mathbb{E}\left(\text{etr}\left(j\mathbf{BT}^T\mathbf{AX}\right)\right)\text{etr}\left(j\mathbf{T}^T\mathbf{C}\right) \\
&= \phi_{\mathbf{X}}\left(\mathbf{A}^T\mathbf{TB}^T\right)\text{etr}\left(j\mathbf{T}^T\mathbf{C}\right) \\
&= \text{etr}\left(j\mathbf{BT}^T\mathbf{AM}\right)\psi\left(\text{Tr}\left(\mathbf{BT}^T\mathbf{A\Sigma A}^T\mathbf{TB}^T\mathbf{\Phi}\right)\right)\text{etr}\left(j\mathbf{T}^T\mathbf{C}\right) \\
&= \text{etr}\left(j\mathbf{T}^T(\mathbf{AMB} + \mathbf{C})\right)\psi\left(\text{Tr}\left(\mathbf{T}^T\left(\mathbf{A\Sigma A}^T\right)\mathbf{T}\left(\mathbf{B}^T\mathbf{\Phi}\mathbf{B}\right)\right)\right)
\end{aligned}$$

This is the characteristic function of $E_{q,m}\left(\mathbf{AMB} + \mathbf{C}, \left(\mathbf{A}^T\mathbf{\Sigma}\mathbf{A}\right) \otimes \left(\mathbf{B}^T\mathbf{\Phi}\mathbf{B}\right), \Psi\right)$. □

Example 8.12.1 (massive MIMO) Consider an MIMO channel

$$\mathbf{y} = \mathbf{Hx} + \mathbf{w}$$

where $\mathbf{H}$ is the channel transfer function and $\mathbf{w}$ the Gaussian random vector. If we consider repeated measurements, we obtain the random matrix model

$$\mathbf{Y} = \mathbf{HX} + \mathbf{W}$$

The above model is relevant. □

Let $\mathbf{X} \sim E_{p,n}(\mathbf{M}, \mathbf{\Sigma} \otimes \mathbf{\Phi}, \Psi)$, and let $\mathbf{\Sigma} = \mathbf{AA}^T$, and $\mathbf{\Phi} = \mathbf{B}^T\mathbf{B}$ be rank factorization of $\mathbf{\Sigma}$ and $\mathbf{\Phi}$. That is, $\mathbf{A}$ is $p \times p_1$ and $\mathbf{B}$ is $n \times n_1$, where $p_1 = \text{rank}(\mathbf{\Sigma}), n_1 = \text{rank}(\mathbf{\Phi})$. Then,

$$\mathbf{A}^{\dagger}(\mathbf{X} - \mathbf{M})\mathbf{B}^{\dagger} \sim E_{p,n}\left(\mathbf{0}, \mathbf{I}_{p_1} \otimes \mathbf{I}_{n_1}, \Psi\right)$$

where the dagger of $\mathbf{A}^{\dagger}$ represents the generalized inverse of $\mathbf{A}$, i.e., $\mathbf{AA}^{\dagger}\mathbf{A} = \mathbf{A}$. Conversely, if $\mathbf{Y} \sim E_{p,n}\left(\mathbf{0}, \mathbf{I}_{p_1} \otimes \mathbf{I}_{n_1}, \Psi\right)$, then

$$\mathbf{AYB}^T + \mathbf{M} \sim E_{p,n}(\mathbf{M}, \mathbf{\Sigma} \otimes \mathbf{\Phi}, \Psi)$$

where $\mathbf{\Sigma} = \mathbf{AA}^T$, and $\mathbf{\Phi} = \mathbf{B}^T\mathbf{B}$

The distribution of $E_p\left(\mathbf{0}, \mathbf{I}_p, \Psi\right)$ is called spherical distribution.

A consequence of the definition of the m.e.c. distribution is that if $\mathbf{X}$ has m.e.c. distribution, then its transpose $\mathbf{X}^T$ has also m.e.c. distribution. Let $\mathbf{X} \sim E_{p,n}(\mathbf{M}, \mathbf{\Sigma} \otimes \mathbf{\Phi}, \Psi)$, then, $\mathbf{X}^T \sim E_{p,n}\left(\mathbf{M}^T, \mathbf{\Phi} \otimes \mathbf{\Sigma}, \Psi\right)$.

The question arises whether the parameters in the definition of a m.e.c. distribution are uniquely defined. The answer is that they are not.

An important subclass of the class of the m.e.c. distribution is the class of matrix variate normal (or Gaussian) distributions. The $p \times n$ random matrix $\mathbf{X}$ is said to have a matrix variate normal distribution if its characteristic function has the form

$$\phi_{\mathbf{X}}(\mathbf{T}) = \text{etr}\left(j\mathbf{T}^T\mathbf{M}\right) \text{etr}\left(-\frac{1}{2}\mathbf{T}^T\mathbf{\Sigma}\mathbf{T}\mathbf{\Phi}\right)$$

with $\mathbf{T}: p \times n, \mathbf{M}: p \times n, \mathbf{\Sigma}: p \times p, \mathbf{\Phi}: n \times n, \mathbf{\Sigma} \geqslant 0, \mathbf{\Phi} \geqslant 0$. This distribution is denoted by $\mathcal{N}_{p,n}(\mathbf{M}, \mathbf{\Sigma} \otimes \mathbf{\Phi})$.

The next theorem shows that the matrix variate normal distribution can be used to represent samples taken from multivariate normal distributions. Let $\mathbf{X} \sim \mathcal{N}_{p,n}\left(\mathbf{me}_n^T, \mathbf{\Sigma} \otimes \mathbf{I}_n\right)$, where $\mathbf{m} \in \mathbb{C}^p$. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be columns of $\mathbf{X}$. Then, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are independent identically distributed random vectors with common distribution $\mathcal{N}_p(\mathbf{m}, \mathbf{\Sigma})$. Here $\mathbf{e}_n$ is the n-dimensional vector whose elements are 1s; that is, $\mathbf{e}_n = (1, 1, \ldots, 1)^T$ real matrix.

8.13 Hypothesis Testing for Matrix Elliptically Contoured Distributions

8.13.1 General Results

If $\mathbf{X} \sim E_{p,n}(\mathbf{M}, \mathbf{\Sigma} \otimes \mathbf{\Phi}, \Psi)$ defines an absolutely continuous elliptically contoured distribution, $\mathbf{\Sigma}$ and $\mathbf{\Phi}$ must be positive definite. The probability density function (p.d.f.) of an m.e.c. distribution is of a special form as the following theorem shows.

Let $\mathbf{X}$ be a $p \times n$ dimensional random matrix whose distribution is absolutely continuous. Then, $\mathbf{X} \sim E_{p,n}(\mathbf{M}, \mathbf{\Sigma} \otimes \mathbf{\Phi}, \Psi)$, if and only if the p.d.f. of $\mathbf{X}$ has the form

$$f(\mathbf{X}) = (\det \mathbf{\Sigma})^{-n/2} (\det \mathbf{\Phi})^{-p/2} h\left(\text{Tr}\left((\mathbf{X} - \mathbf{M})^T \mathbf{\Sigma}^{-1} (\mathbf{X} - \mathbf{M}) \mathbf{\Phi}^{-1}\right)\right)$$

where h and Ψ determine each other for the specified p and n.

If $\mathbf{X} \sim E_{p,n}(\mathbf{M}, \mathbf{\Sigma} \otimes \mathbf{\Phi}, \Psi)$, (a) if $\mathbf{X}$ has a finite first moment, then $\mathbb{E}(\mathbf{X}) = \mathbf{M}$; (b) if $\mathbf{X}$ has a finite second moment, then $\text{Cov}(\mathbf{X}) = c\mathbf{\Sigma} \otimes \mathbf{\Phi}$, where $c = -2\psi'(0)$, with $\psi'(t)$ the first derivative.

We can give the stochastic representation of a m.e.c. distribution. Let $\mathbf{X}$ be a $p \times n$ random matrix. Let $\mathbf{M}: p \times n, \mathbf{\Sigma}: p \times p, \mathbf{\Phi}: n \times n$ be constant matrices, $\mathbf{\Sigma} \geqslant 0, \mathbf{\Phi} \geqslant 0$, $\text{rank}(\mathbf{\Sigma}) = p_1, \text{rank}(\mathbf{\Phi}) = n_1$. Then

$$\mathbf{X} \sim E_{p,n}(\mathbf{M}, \mathbf{\Sigma} \otimes \mathbf{\Phi}, \Psi)$$

if and only if

$$\mathbf{X} \approx \mathbf{M} + r\mathbf{AUB}^T$$

where $\mathbf{U}$ is $p_1 \times n_1$ and vec $\left(\mathbf{U}^T\right)$ is uniformly distributed on $S_{p_1 n_1}$, r is a non-negative random variable, r and $\mathbf{U}$ are independent, $\mathbf{\Sigma} = \mathbf{A}\mathbf{A}^T$, and $\mathbf{\Phi} = \mathbf{B}\mathbf{B}^T$ are rank factorizations of $\mathbf{\Sigma}$ and $\mathbf{\Phi}$. Moreover

$$\psi(u) = \int_0^\infty \Omega_{p_1 n_1}\left(r^2 u\right) dF(r), u \geqslant 0$$

where $\Omega_{p_1 n_1}\left(\mathbf{t}^T\mathbf{t}\right), \mathbf{t} \in \mathbb{R}^{p_1 n_1}$ denotes the characteristic function of vec $\left(\mathbf{U}^T\right)$, and $F(r)$ denotes the distribution function of r. The expression $\mathbf{M} + r\mathbf{A}\mathbf{U}\mathbf{B}^T$ is called the stochastic representation of $\mathbf{X}$. Let us denote the unit sphere in $\mathbb{R}^k$ by S_k

$$S_k = \left\{\mathbf{x} | \mathbf{x} \in \mathbb{R}^k; \mathbf{x}^T\mathbf{x} = 1\right\}$$

Let $\mathbf{X} \sim E_{p,n}(\mathbf{M}, \mathbf{\Sigma} \otimes \mathbf{\Phi}, \Psi)$, and $r\mathbf{A}\mathbf{U}\mathbf{B}^T$ be a stochastic representation of $\mathbf{X}$. Assume $\mathbf{X}$ is absolutely continuous and has the p.d.f.

$$f(\mathbf{X}) = (\det \mathbf{\Sigma})^{-n/2} (\det \mathbf{\Phi})^{-p/2} h\left(\mathrm{Tr}\left(\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X}\mathbf{\Phi}^{-1}\right)\right)$$

Then, r is also absolutely continuous and has the p.d.f.

$$g(r) = \frac{2\pi^{pn/2}}{\Gamma\left(\frac{pn}{2}\right)} r^{pn-1} h\left(r^2\right), \quad r \geqslant 0$$

The stochastic representation is a major tool in the study of m.e.c. distributions.

Theorem 8.13.1 Let $\mathbf{X} \sim E_{p,n}(\mathbf{M}, \mathbf{\Sigma} \otimes \mathbf{\Phi}, \Psi)$, with p.d.f.

$$f(\mathbf{X}) = (\det \mathbf{\Sigma})^{-n/2} (\det \mathbf{\Phi})^{-p/2} h\left(\mathrm{Tr}\left((\mathbf{X} - \mathbf{M})^T \mathbf{\Sigma}^{-1} (\mathbf{X} - \mathbf{M}) \mathbf{\Phi}^{-1}\right)\right)$$

where $h(z)$ is monotone decreasing on $[0, \infty)$. Suppose $h, \mathbf{\Sigma}$ and $\mathbf{\Phi}$ are known and we want to find the maximum likelihood estimate (MLE) of $\mathbf{M}$ (say $\hat{\mathbf{M}}$), based on a single observation $\mathbf{X}$. Then,

(a) $\hat{\mathbf{M}} = \mathbf{X}$
(b) If $\mathbf{M} = \boldsymbol{\mu}\mathbf{v}^T$, where $\boldsymbol{\mu}$ is p-dimensional, $\mathbf{v}$ is n-dimensional vector and $\mathbf{v} \neq \mathbf{0}$ is known, the MLE of $\boldsymbol{\mu}$ is $\hat{\boldsymbol{\mu}} = \mathbf{X}\frac{\mathbf{\Phi}^{-1}\mathbf{v}}{\mathbf{v}^T\mathbf{\Phi}^{-1}\mathbf{v}}$, and
(c) if $\mathbf{M}$ is of the form $\mathbf{M} = \boldsymbol{\mu}\mathbf{e}_n^T$, the MLE of $\boldsymbol{\mu}$ is $\boldsymbol{\mu} = \mathbf{X}\frac{\mathbf{\Phi}^{-1}\mathbf{e}_n}{\mathbf{e}_n^T\mathbf{\Phi}^{-1}\mathbf{e}_n}$

Now we are ready to state our result on the likelihood ratio test (LRT) statistic. Assume we have an observation $\mathbf{X}$ from the distribution $E_{p,n}(\mathbf{M}, \mathbf{\Sigma} \otimes \mathbf{\Phi}, \Psi)$, and we want to test

$$\mathcal{H}_0 : (\mathbf{M}, \mathbf{\Sigma} \otimes \mathbf{\Phi}) \in \omega \quad \text{against} \quad \mathcal{H}_1 : (\mathbf{M}, \mathbf{\Sigma} \otimes \mathbf{\Phi}) \in \Omega - \omega \tag{8.235}$$

where $\omega \subset \Omega$. Suppose Ω and ω have the properties that if $\mathbf{Q} \in \mathbb{R}^{p \times n}, \mathbf{S} \in \mathbb{R}^{pn \times pn}$, then $(\mathbf{Q}, \mathbf{S}) \in \Omega$ implies $(\mathbf{Q}, c\mathbf{S}) \in \Omega$, and $(\mathbf{Q}, \mathbf{S}) \in \omega$ implies $(\mathbf{Q}, c\mathbf{S}) \in \omega$ for any positive scalar c. Moreover, let $\mathbf{X}$ have the p.d.f.

$$f(\mathbf{X}) = (\det \mathbf{\Sigma})^{-n/2} (\det \mathbf{\Phi})^{-p/2} h\left(\mathrm{Tr}\left((\mathbf{X} - \mathbf{M})^T \mathbf{\Sigma}^{-1} (\mathbf{X} - \mathbf{M}) \mathbf{\Phi}^{-1}\right)\right)$$

where $l(z) = z^{pn/2} h(z) (z \geqslant 0)$ has a finite maximum at $z = z_h > 0$.

Furthermore, suppose that under the assumption $\mathcal{H}_1$ that $\mathbf{X} \sim \mathcal{N}_{p,n}(\mathbf{M}, \mathbf{\Sigma} \otimes \mathbf{\Phi})$, $(\mathbf{M}, \mathbf{\Sigma} \otimes \mathbf{\Phi}) \in \Omega$, the MLE's of $\mathbf{M}$ and $\mathbf{\Sigma} \otimes \mathbf{\Phi}$ are $\mathbf{M}^*$ and $\mathbf{\Sigma} \otimes \mathbf{\Phi}^*$, which are unique and $\mathbb{P}\left((\mathbf{\Sigma} \otimes \mathbf{\Phi})^* > 0\right) = 1$. Assume also that under the assumption $\mathcal{H}_0$

$\mathbf{X} \sim \mathcal{N}_{p,n}(\mathbf{M}, \mathbf{\Sigma} \otimes \mathbf{\Phi})$, $(\mathbf{M}, \mathbf{\Sigma} \otimes \mathbf{\Phi}) \in \omega$, the MLE's are $\mathbf{M}$ and $\mathbf{\Sigma} \otimes \mathbf{\Phi}$ are $\mathbf{M}_0^*$ and $\mathbf{\Sigma} \otimes \mathbf{\Phi}_0^*$, which are unique and $\mathbb{P}\left((\mathbf{\Sigma} \otimes \mathbf{\Phi})_0^* > 0\right) = 1$.

Then, the likelihood ratio test (LRT) statistic for testing (8.235) under the assumption that $\mathbf{X} \sim E_{p,n}(\mathbf{M}, \mathbf{\Sigma} \otimes \mathbf{\Phi}, \Psi)$, is the **same** as under the assumption that $\mathbf{X} \sim \mathcal{N}_{p,n}(\mathbf{M}, \mathbf{\Sigma} \otimes \mathbf{\Phi})$, namely

$$\frac{\det(\mathbf{\Sigma} \otimes \mathbf{\Phi})^*}{\det(\mathbf{\Sigma} \otimes \mathbf{\Phi})_0^*}$$

8.13.2 Two Models

Now, we describe the parameter spaces in which we want to study hypothesis testing problems.

Model I

Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be p-dimensional random vectors, such that $n > p$ and $\mathbf{x}_i \sim E_p(\boldsymbol{\mu}, \mathbf{\Sigma}, \psi), i = 1, \ldots, n$. Moreover, assume that $\mathbf{x}_i, i = 1, \ldots, n$ are uncorrelated and their joint distribution is elliptically contoured and absolutely continuous. This model can be expressed as

$$\mathbf{X} \sim E_{p,n}\left(\boldsymbol{\mu}\mathbf{e}_n^T, \mathbf{\Sigma} \otimes \mathbf{I}_n, \psi\right) \tag{8.236}$$

where $\mathbf{X} = \left(\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_n\right)$. Then the joint p.d.f. of $\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_n$ can be written as

$$f(\mathbf{X}) = \frac{1}{(\det \mathbf{\Sigma})^n} h\left(\sum_{i=1}^{n} \left(\mathbf{x}_i - \boldsymbol{\mu}\right)^T \mathbf{\Sigma}^{-1} \left(\mathbf{x}_i - \boldsymbol{\mu}\right)\right) \tag{8.237}$$

Assume $l(z) = z^{pn/2} h(z), z \geqslant 0$ has a finite maximum at $z = z_h > 0$. Define

$$\overline{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \quad \text{and} \quad \mathbf{A} = \sum_{i=1}^{n} \left(\mathbf{x}_i - \overline{\mathbf{x}}\right) \left(\mathbf{x}_i - \overline{\mathbf{x}}\right)^T$$

Then $\overline{\mathbf{x}} = \frac{1}{n}\mathbf{X}\mathbf{e}_n, \mathbf{A} = \mathbf{X}\left(\mathbf{I}_n - \frac{1}{n}\mathbf{e}_n\mathbf{e}_n^T\right)\mathbf{X}^T$, and the statistic $T(\mathbf{X}) = \left(\overline{\mathbf{x}}, \mathbf{A}\right)$ is sufficient for $(\boldsymbol{\mu}, \mathbf{\Sigma})$.

If $\psi(z) = \exp\left(-\frac{z}{2}\right)$, then $\mathbf{X} \sim \mathcal{N}_{p,n}\left(\boldsymbol{\mu}\mathbf{e}_n^T, \mathbf{\Sigma} \otimes \mathbf{I}_n, \psi\right)$. In this case, $\mathbf{x}_i$ are independent, and identically distributed random vectors each with distribution $\mathcal{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$. Inference for this structure has been extensively studied in Anderson [371].

Model II

Let $\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \ldots, \mathbf{x}_{n_i}^{(i)}$ be p-dimensional random vectors, such that $n_i > p, i = 1, \ldots, q$, and $\mathbf{x}_j^{(i)} \sim E_p\left(\mu_i, \mathbf{\Sigma}_i, \psi\right), j = 1, \ldots, n_i, i = 1, \ldots, q$. Moreover, assume that $\mathbf{x}_j^{(i)}, i = 1, \ldots, q, j = 1, \ldots, n_i$ are uncorrelated and their joint distribution is also elliptically contoured and absolutely continuous. This model can be expressed as

$$\mathbf{x} \sim E_{p,n}\left(\begin{pmatrix} \mathbf{e}_{n_1} \otimes \boldsymbol{\mu}_1 \\ \mathbf{e}_{n_1} \otimes \boldsymbol{\mu}_2 \\ \vdots \\ \mathbf{e}_{n_1} \otimes \boldsymbol{\mu}_q \end{pmatrix}, \begin{pmatrix} \mathbf{I}_{n_1} \otimes \mathbf{\Sigma}_1 \\ \mathbf{I}_{n_2} \otimes \mathbf{\Sigma}_2 \\ \vdots \\ \mathbf{I}_{n_q} \otimes \mathbf{\Sigma}_q \end{pmatrix}, \psi\right) \tag{8.238}$$

where $n = \sum_{i=1}^{q} n_i$ and

$$\mathbf{x} = \left[\mathbf{x}_1^{(1)}, \ldots, \mathbf{x}_{n_1}^{(1)}, \mathbf{x}_1^{(2)}, \ldots, \mathbf{x}_{n_2}^{(2)}, \mathbf{x}_1^{(q)}, \ldots, \mathbf{x}_{n_q}^{(q)}\right]^T$$

Then, the joint p.d.f. of $\mathbf{x}_j^{(i)}, i = 1, \ldots, q, j = 1, \ldots, n_i$ can be written as

$$f(\mathbf{x}) = \frac{1}{\prod_{i=1}^{q} \left(\det \boldsymbol{\Sigma}_i\right)^{n_i}} h\left(\sum_{i=1}^{q}\sum_{j=1}^{n_i} \left(\mathbf{x}_j^{(i)} - \boldsymbol{\mu}_i\right)^T \boldsymbol{\Sigma}_i^{-1} \left(\mathbf{x}_j^{(i)} - \boldsymbol{\mu}_i\right)\right) \tag{8.239}$$

Assume $l(z) = z^{pn/2} h(z), z \geqslant 0$ has a finite maximum at $z = z_h > 0$. Define

$$\overline{\mathbf{x}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbf{x}_j^{(i)}, \qquad \mathbf{A}_i = \sum_{j=1}^{n_i} \left(\mathbf{x}_j^{(i)} - \overline{\mathbf{x}}_i\right)\left(\mathbf{x}_j^{(i)} - \overline{\mathbf{x}}_i\right)^T$$

and $\mathbf{A} = \sum_{i=1}^{q} \mathbf{A}_i$. Also let $\overline{\mathbf{x}} = \sum_{i=1}^{q}\sum_{j=1}^{n_i} \mathbf{x}_j^{(i)}$ and $\mathbf{B} = \sum_{i=1}^{q}\sum_{j=1}^{n_i} \left(\mathbf{x}_j^{(i)} - \overline{\mathbf{x}}\right)\left(\mathbf{x}_j^{(i)} - \overline{\mathbf{x}}\right)^T$. Then, we get

$$\begin{aligned}
&\sum_{j=1}^{n_i} \left(\mathbf{x}_j^{(i)} - \boldsymbol{\mu}_i\right)^T \boldsymbol{\Sigma}_i^{-1} \left(\mathbf{x}_j^{(i)} - \boldsymbol{\mu}_i\right) \\
&\quad = \operatorname{Tr}\left(\sum_{j=1}^{n_i} \left(\mathbf{x}_j^{(i)} - \boldsymbol{\mu}_i\right)^T \boldsymbol{\Sigma}_i^{-1} \left(\mathbf{x}_j^{(i)} - \boldsymbol{\mu}_i\right)\right) \\
&\quad = \operatorname{Tr}\left(\boldsymbol{\Sigma}_i^{-1} \sum_{j=1}^{n_i} \left(\mathbf{x}_j^{(i)} - \boldsymbol{\mu}_i\right)^T \left(\mathbf{x}_j^{(i)} - \boldsymbol{\mu}_i\right)\right) \\
&\quad = \operatorname{Tr}\left(\boldsymbol{\Sigma}_i^{-1}\left(\sum_{j=1}^{n_i} \left(\mathbf{x}_j^{(i)} - \overline{\mathbf{x}}_i\right)^T \left(\mathbf{x}_j^{(i)} - \overline{\mathbf{x}}_i\right) + n\left(\overline{\mathbf{x}}_i - \boldsymbol{\mu}_i\right)\left(\overline{\mathbf{x}}_i - \boldsymbol{\mu}_i\right)^T\right)\right) \\
&\quad = \operatorname{Tr}\left(\boldsymbol{\Sigma}_i^{-1}\left(\mathbf{A}_i + n\left(\overline{\mathbf{x}}_i - \boldsymbol{\mu}_i\right)\left(\overline{\mathbf{x}}_i - \boldsymbol{\mu}_i\right)^T\right)\right)
\end{aligned}$$

Thus,

$$f(\mathbf{x}) = \frac{1}{\prod_{i=1}^{q} \left(\det \boldsymbol{\Sigma}_i\right)^{n_i}} h\left(\sum_{i=1}^{q} \operatorname{Tr}\left(\boldsymbol{\Sigma}_i^{-1}\left(\mathbf{A}_i + n\left(\overline{\mathbf{x}}_i - \boldsymbol{\mu}_i\right)\left(\overline{\mathbf{x}}_i - \boldsymbol{\mu}_i\right)^T\right)\right)\right)$$

hence the statistic $\left(\overline{\mathbf{x}}^{(1)}, \ldots, \overline{\mathbf{x}}^{(q)}, \mathbf{A}_1, \ldots, \mathbf{A}_q\right)$ is sufficient for $\left(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_q\right)$

If $\psi(z) = \exp\left(-\frac{z}{2}\right)$, then

$$\mathbf{x} \sim \mathcal{N}_{p,n}\left(\begin{pmatrix} \mathbf{e}_{n_1} \otimes \boldsymbol{\mu}_1 \\ \mathbf{e}_{n_1} \otimes \boldsymbol{\mu}_2 \\ \vdots \\ \mathbf{e}_{n_1} \otimes \boldsymbol{\mu}_q \end{pmatrix}, \begin{pmatrix} \mathbf{I}_{n_1} \otimes \Sigma_1 \\ \mathbf{I}_{n_2} \otimes \Sigma_2 \\ \vdots \\ \mathbf{I}_{n_q} \otimes \Sigma_q \end{pmatrix}\right)$$

In this case, $\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \ldots, \mathbf{x}_{n_i}^{(i)}$ are independent, and identically distributed random variables each with distribution $\mathcal{N}_p\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right), i = 1, \ldots, q$. Moreover, $\mathbf{x}_j^{(i)}, i = 1, \ldots, q, j = 1, \ldots, n_i$ are jointly independent. Inference for this structure has been studied in [371].

A special case of Model II is when $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_q = \boldsymbol{\Sigma}$. Then the model can also be expressed as

$$\mathbf{X} \sim E_{p,n}\left(\left(\boldsymbol{\mu}_1 \mathbf{e}_{n_1}^T, \boldsymbol{\mu}_2 \mathbf{e}_{n_2}^T, \ldots, \boldsymbol{\mu}_q \mathbf{e}_{n_q}^T\right), \boldsymbol{\Sigma} \otimes \mathbf{I}_n, \psi\right)$$

where $n = \sum_{i=1}^{q} n_i$ and

$$\mathbf{x} = \left(\mathbf{x}_1^{(1)}, \ldots, \mathbf{x}_{n_1}^{(1)}, \mathbf{x}_1^{(2)}, \ldots, \mathbf{x}_{n_2}^{(2)}, \mathbf{x}_1^{(q)}, \ldots, \mathbf{x}_{n_q}^{(q)}\right)$$

This leads to the same joint p.d.f. of $\mathbf{x}_j^{(i)}, i = 1, \ldots, q, j = 1, \ldots, n_i$ as (8.238); that is

$$f(\mathbf{x}) = \frac{1}{(\det \Sigma)^n} h\left(\sum_{i=1}^{q} \sum_{j=1}^{n_i} \left(\mathbf{x}_j^{(i)} - \boldsymbol{\mu}_i\right)^T \Sigma^{-1} \left(\mathbf{x}_j^{(i)} - \boldsymbol{\mu}_i\right)\right)$$

8.13.3 Testing Criteria

We only list criteria that are of interest to us.

Testing that a Covariance Matrix is Equal to a Given Matrix

In Model I (Section 8.13.2), assume that h is decreasing. We want to test

$$\mathcal{H}_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 \quad \text{against} \quad \mathcal{H}_1 : \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}_0 \tag{8.240}$$

We assume that $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown and $\boldsymbol{\Sigma}_0 > \mathbf{0}$ is given. We can show that (8.240) is equivalent to testing

$$\mathcal{H}_0 : \boldsymbol{\Sigma} = \mathbf{I}_p \quad \text{against} \quad \mathcal{H}_1 : \boldsymbol{\Sigma} \neq \mathbf{I}_p \tag{8.241}$$

Theorem 8.13.2 The LRT statistic for the problem (8.240) is

$$\tau = \left(\det\left(\boldsymbol{\Sigma}_0^{-1}\mathbf{A}\right)\right)^{n/2} h\left(\operatorname{Tr}\left(\boldsymbol{\Sigma}_0^{-1}\mathbf{A}\right)\right)$$

The critical region at level α is

$$\tau \leqslant \tau_\psi(\alpha)$$

where $\tau_\psi(\alpha)$ depends on ψ, but not on $\boldsymbol{\Sigma}_0$. The null distribution of τ does not depend on $\boldsymbol{\Sigma}_0$.

Testing that a Covariance Matrix is Proportional to a Given Matrix

In Model I (Section 8.13.2), we want to test

$$\mathcal{H}_0 : \boldsymbol{\Sigma} = \sigma^2 \boldsymbol{\Sigma}_0 \quad \text{against} \quad \mathcal{H}_1 : \boldsymbol{\Sigma} \neq \sigma^2 \boldsymbol{\Sigma}_0 \tag{8.242}$$

where $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \sigma^2$ are unknown, and $\sigma^2 > 0$ is a scalar, and $\boldsymbol{\Sigma}_0 > \mathbf{0}$ is given. Problem (8.242) remains invariant under group G, where G is generated by the linear transformations

- $g(\mathbf{X}) = c\mathbf{X}, \quad c > 0$ scalar and
- $g(\mathbf{X}) = \mathbf{X} + \mathbf{v}\mathbf{e}_n^T, \quad \mathbf{v}$ is $a\ p$ - dimensional vector.

It is easy to show that (8.242) is equivalent to testing

$$\mathcal{H}_0 : \mathbf{\Sigma} = \sigma^2 \mathbf{I}_p \quad \text{against} \quad \mathcal{H}_1 : \mathbf{\Sigma} \neq \sigma^2 \mathbf{I}_p \tag{8.243}$$

Theorem 8.13.3 The LRT statistic for the problem (8.242) is

$$\tau^{2/n} = \frac{\det\left(\mathbf{\Sigma}_0^{-1}\mathbf{A}\right)}{\left(\mathrm{Tr}\left(\frac{1}{p}\mathbf{\Sigma}_0^{-1}\mathbf{A}\right)\right)^p}$$

The critical region at level α is

$$\tau \leqslant \tau_\psi(\alpha)$$

where $\tau_\psi(\alpha)$ is the same as in the normal (Gaussian) case and does not depend on $\mathbf{\Sigma}_0$.

The distribution of τ is the same as in the normal (Gaussian) case. The null distribution of τ does not depend on $\mathbf{\Sigma}_0$. τ is an invariant of the sufficient statistic under the group G.

Testing Equality of Covariance Matrices

In Model II (Section 8.13.2), we want to test

$$\begin{aligned} &\mathcal{H}_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \cdots = \mathbf{\Sigma}_q \\ &\mathcal{H}_1 : \text{there exist } 1 \leqslant j \leqslant k \leqslant q, \text{ such that } \mathbf{\Sigma}_j \neq \mathbf{\Sigma}_k \end{aligned} \tag{8.244}$$

where $\boldsymbol{\mu}_i$ and $\mathbf{\Sigma}_i$, $i = 1, 2, \ldots, q$ are unknown. Problem (8.244) remains invariant under group G, where G is generated by the linear transformations

- $g(\mathbf{X}) = \left(\mathbf{I}_n \otimes \mathbf{C}\right)\mathbf{X}, \quad$ where $\mathbf{C}$ is $p \times p$ nonsingular matrix;
- $g(\mathbf{X}) = \mathbf{X} - \begin{pmatrix} \mathbf{e}_{n_1} \otimes \mathbf{v}_1 \\ \mathbf{e}_{n_2} \otimes \mathbf{v}_2 \\ \vdots \\ \mathbf{e}_{n_q} \otimes \mathbf{v}_q \end{pmatrix}, \quad \mathbf{v}_i$ is $p-$dimensional vector, $i = 1, 2, \ldots, q$.

Theorem 8.13.4 The LRT statistic for the problem (8.244) is

$$\tau = \frac{\prod_{i=1}^{q}\left[\det\left(\mathbf{A}_i\right)\right]^{n_i/2}\prod_{i=1}^{q}\left(n_i\right)^{pn_i/2}}{\left(\det\mathbf{A}\right)^{n/2} n^{pn/2}}$$

The critical region at level α is

$$\tau \leqslant \tau_\psi(\alpha)$$

where $\tau_\psi(\alpha)$ is the same as in the normal (Gaussian) case. The distribution of τ is the same as in the normal (Gaussian) case. τ is an invariant of the sufficient statistic under the group G.

Testing Equality of Means and Covariance Matrices

In Model II (Section 8.13.2), we want to test

$$\begin{aligned} &\mathcal{H}_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_q \text{ and } \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_q \\ &\mathcal{H}_1 : \text{there exist } 1 \leqslant j \leqslant k \leqslant q, \text{ such that } \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2 \text{ or } \boldsymbol{\Sigma}_j \neq \boldsymbol{\Sigma}_k \end{aligned} \tag{8.245}$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$, $i = 1, 2, \ldots, q$ are unknown. Problem (8.245) remains invariant under group G, where G is generated by the linear transformations

- $g(\mathbf{X}) = \left(\mathbf{I}_n \otimes \mathbf{C}\right)\mathbf{X}$, where $\mathbf{C}$ is $p \times p$ nonsingular matrix;
- $g(\mathbf{X}) = \mathbf{X} - \mathbf{e}_n \otimes \mathbf{v}$, where $\mathbf{v}$ is p–dimensional vector.

Theorem 8.13.5 The LRT statistic for the problem (8.245) is

$$\tau = \frac{\prod_{i=1}^{q} \left[\det\left(\mathbf{A}_i\right)\right]^{n_i/2}}{\left(\det \mathbf{B}\right)^{n/2}}$$

The critical region at level α is

$$\tau \leqslant \tau_{\psi}(\alpha)$$

where $\tau_{\psi}(\alpha)$ is the same as in the normal (Gaussian) case. The distribution of τ is the same as in the normal (Gaussian) case. τ is an invariant of the sufficient statistic under the group G.

Bibliographical Remarks

For more details on eigenvalue variance bounds, we refer to Dallaporta's PhD dissertation [435], from which we take the liberty of freely drawing material for Section 8.3. We want a theory of random matrices, which is valid for arbitrary sizes of matrices [442, 518, 519]. In particular, the book [40] gives an relatively comprehensive survey of the results in this context.

Section 8.4 is taken from [448, 450] with minor notation changes. There is a large literature on this subject. See e.g., Qiu and Wicks [40] for some recent results.

Regarding Section 8.3.2, Wasserstein distances is treated in [443], from which we draw some material.

In Section 8.8, we study how large random matrices will be perturbed by finite rank perturbations, drawing material from [367].

In Section 6.18, we take material from [340]. The discovery of outliers in Figure 6.39 was made in July 2013 during the writing of Section 6.18: see Figure 6.39 and the associated MATLAB code. The outlier part of Section 6.18 is taken from [368]. In Section 6.15, we draw material from [328, 329], [340] and [368].

In Section 6.16, outliers in the spectrum of i.i.d matrices with bounded rank perturbations are considered, taking material from [345]. Another good reference is "around the circular law" [328].

In Section 6.17, we took the liberty of freely drawing material from Benaych-George and Rochet (2013) [138].

We have used material from Ledoit and Wolf (2002) [469] and Srivastava (2005) [477] in Section 8.9. The proofs of the results can be found there.

In Section 8.9, two PhD dissertations [499, 520] and a MS thesis [506] are recommended, from which we have taken some material. We also follow [474] for many developments in this section. Section 8.9.4 is taken from [489] with some notation changes. In Section 8.9.2 we drew on material from [486] and [521, p. 587].

The material in Section 3.8 may be found in [191].

Most material in Section 8.9.8 can be found in [507].

Most material in Section 8.9.9 can be found in [522].

Section 8.9.10 is taken from [511, 523].

Section 8.10 is taken from [513].

Covariance matrix estimation with a well-structured eigensystem is desirable. Authors in [381] proposed estimating the covariance matrix through its matrix logarithm based on an approximate log-likelihood function. The PhD dissertation [524] studies two sample inference for high-dimensional data. Instead of relying on the eigenvalue distribution, [525] devises a linear *shrinkage*-based minimum description length (LS-MDL) criterion by utilizing the identity covariance matrix structure of noise subspace components. The eigenvalues obtained from the estimator turn out to be a linear function of the corresponding sample eigenvalues, enabling the LS-MDL criterion to accurately detect the source number without incurring significantly additional computational load.

Couillet and Zio [526] tackle this problem by first demonstrating that successive voltage observations of a line outage can be mathematically modeled by a spiked sample covariance matrix.

Section 8.12 and Section 8.13 are essentially taken from [217] and [433]; [217] is an introduction to the subject and is the first treatment to organize the materials in a unified manner.

Part II

Smart Grid

9

Applications and Requirements of Smart Grid

9.1 History

The shift in the development of transmission grids that has made them more intelligent has been summarized as "smart grid". Other terms such as IntelliGrid, GridWise, FutureGrid, have also been used.

The IntelliGrid program, initiated by the Electric Power Research Institution (EPRI), is to create the technical foundation for a smart power grid that links electricity with communications and computer control to achieve tremendous gains in the enhancement of reliability, capacity, and customer service [527, 528]. Electrical Power Research Institute (EPRI) first proposed the concept of Intelli-Grid around the year 2000, detailing trends in power grids and the solutions to the problems facing the twenty-first century. The Department of Energy (DOE) launched the Grid-Wise project around 2004, with the goal of distribution systems.

Europe used the terminology of Smart Grid. In 2005, the European Smart-Grids Technology Platform [529, 530] was formed, and a report [531] on roadmaps and ideas was issued in 2006. It identified the important features of Europe's electricity networks as being flexible in response to customers' requests, being accessible to network users, being reliable in terms of the security and quality of the power supply, and being economical to provide the best value and energy-efficient management.

A Federal Smart Grid Task Force was established by the US Department of Energy (DoE) under Title XIII of the Energy Independence and Security Act of 2007. In its Grid 2030 vision, the objective is to construct a twenty-first century electric system to provide abundant, affordable, clean, efficient, and reliable electric power anytime, anywhere [532]. The expected achievements, through smart-grid development, will not merely enhance the reliability, efficiency, and security of the nation's electric grid but also contribute to the strategic goal of reducing carbon emissions.

"Release 1.0 NIST Framework and Roadmap for Smart Grid Interoperability" was published in January 2010 [533].

The Smart Grid is expected to be fully functional by 2030. The future electric grid is illustrated in Figure 9.1 [534].

Smart Grid using Big Data Analytics: A Random Matrix Theory Approach, First Edition.
Robert C. Qiu and Paul Antonik.

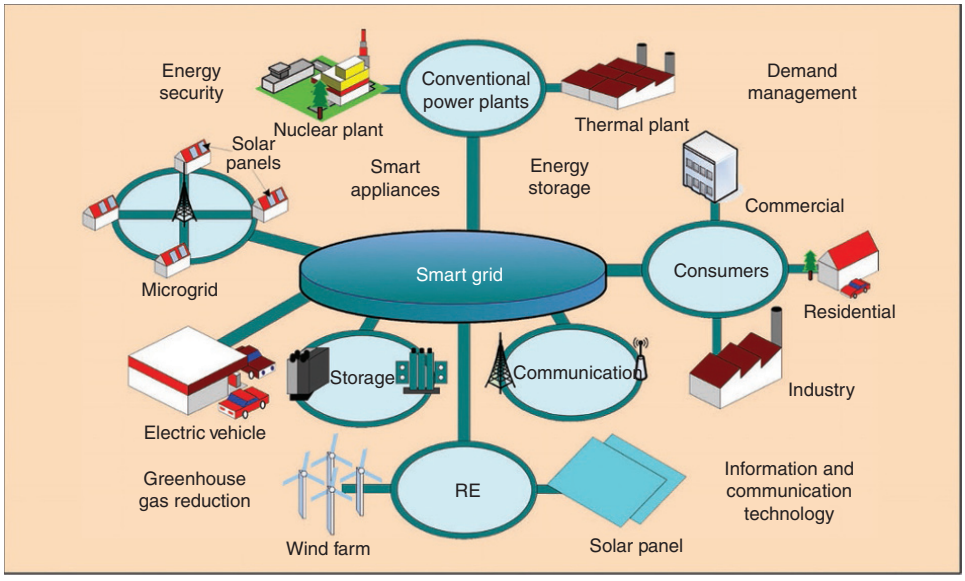


Figure 9.1 The future electric grid. RE: renewable energy. Source: Reproduced from [534] with permission of IEEE.

9.2 Concepts and Vision

The electric grid was cited by the National Academy of Engineering as the supreme engineering achievement of the twentieth century [536]. Industrialization and economic development have historically been associated with man's ability to harness natural energy resources to improve his condition (Figure 9.2). Three dominant factors are affecting the world's future electric systems: governmental policies at both federal and state levels, customer efficiency needs, and new intelligent computer software and hardware technologies. Environmental concerns are also driving the entire energy system to efficiency, conservation, and renewable sources of electricity. Customers are becoming more proactive and are being empowered to engage in energy consumption decisions affecting their day-to-day lives.

The *smart grid* can be defined as an electric system that uses information, two-way, cyber-secure communication technologies, and computational intelligence in an integrated fashion across the entire spectrum of the energy system from the generation to the end points of consumption of the electricity [538]. Smart grid is not a thing but a *vision*. Modernization of the electric grid is a significant long-term undertaking that will span decades. A conceptual model of the envisioned future smart grid is shown in Figure 9.3. Domains (Figure 9.7) in the smart grid conceptual model are defined in Table 9.1. It is envisioned that the electric power grid will move from an electromechanically controlled system to an electronically controlled network in the next two decades [539].

With emerging requirements for renewable portfolio standards, limits on greenhouse gases, and demand response and energy conservation measures, environmental issues have moved to the forefront of the utility business [540]. Figure 9.4 shows projected power generation additions for 2020. Figure 9.5 shows next-generation cost comparison.

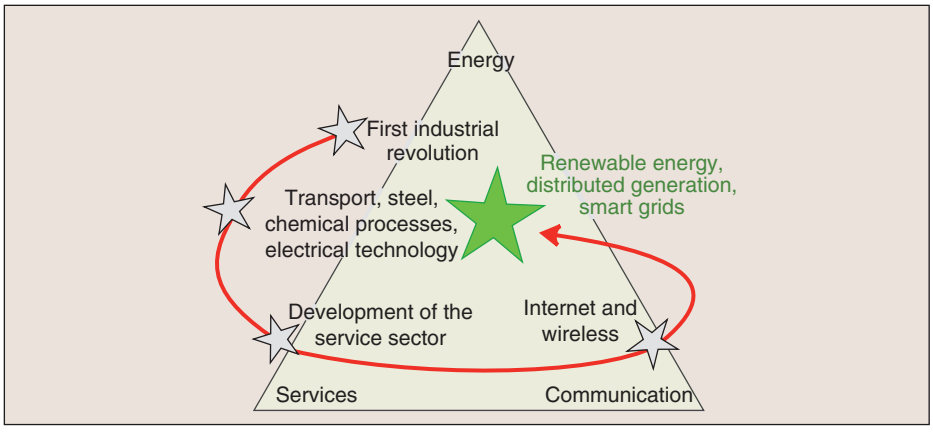


Figure 9.2 Visual history of industrial revolutions: from energy to services and communication and back again to energy. Source: Reproduced from [535] with permission of IEEE.

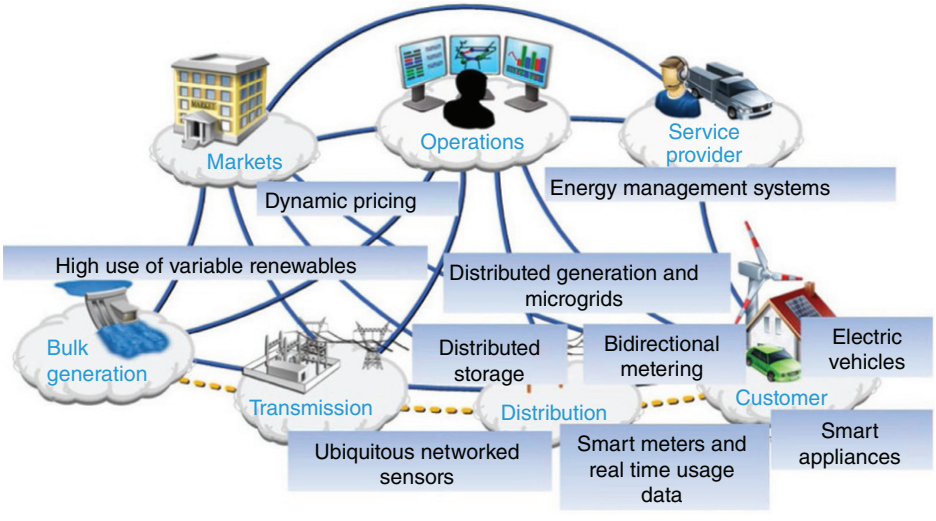


Figure 9.3 Conceptual model of the smart grid. Source: Reproduced from NIST Report. http://www.nist.gov/smartgrid/upload/NISTSP-1108r3.pdf (accessed September 8, 2016).

During 2009, NIST engaged over 1500 stakeholders representing hundreds of organizations in a series of public workshops over a nine-month period to create a high-level architectural model for the smart grid (see Figure 9.6 and Figure 9.7). The result of this work, was published in January 2010 [533].

9.3 Today's Electric Grid

Today's conventional power delivery can be broken into mostly isolated components of generation, transmission, substation, distribution, and customers. The features include [538]:

Table 9.1 Domains in the smart grid conceptual model.

Domain	Actors in the domain
Customers	The end users of electricity. May also generate, store, and manage the use of energy. Traditionally, three customer types are discussed, each with its own domain: residential, commercial/building, and industrial.
Markets	The operators and participants in electricity markets.
Service providers	The organizations providing services to electrical customers and utilities.
Dispatchers	The managers of power dispatch.
Bulk generation	The generators of electricity in bulk quantities. May also store energy for later distribution.
Transmission	The carriers of bulk electricity over long distances.
Distribution	The distributors of electricity to and from customers. May also store and generate electricity.

- centralized sources of power generation;
- unidirectional flow of energy from the sources to the customers;
- passive participation by customers: customer knowledge of electrical energy usage is limited to a monthly bill received, after the fact, at the end of the month;
- real-time monitoring and control is mainly limited to generation and transmission, and only at some utilities does it extend to the distribution system;

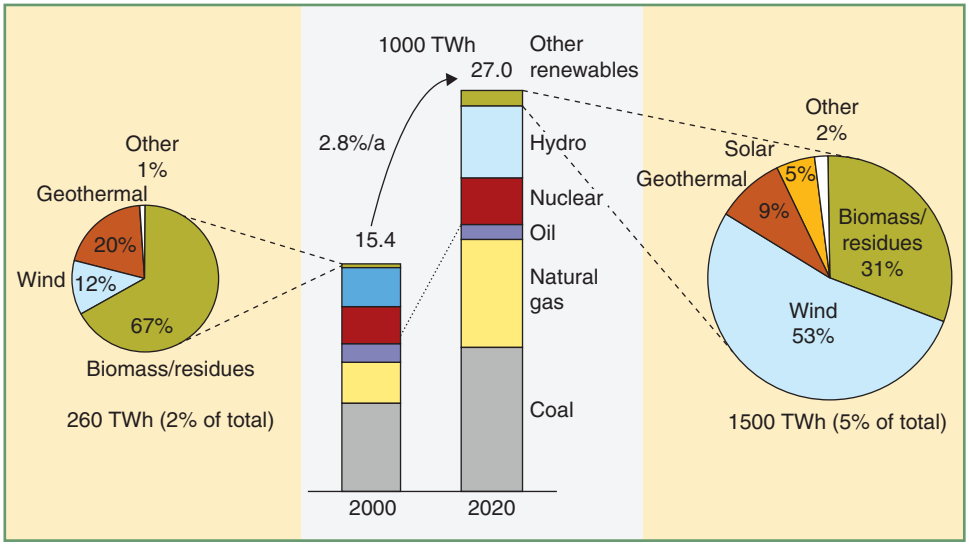


Figure 9.4 Projected power generation additions: 2020. Source: Reproduced from [541] with permission of IEEE.

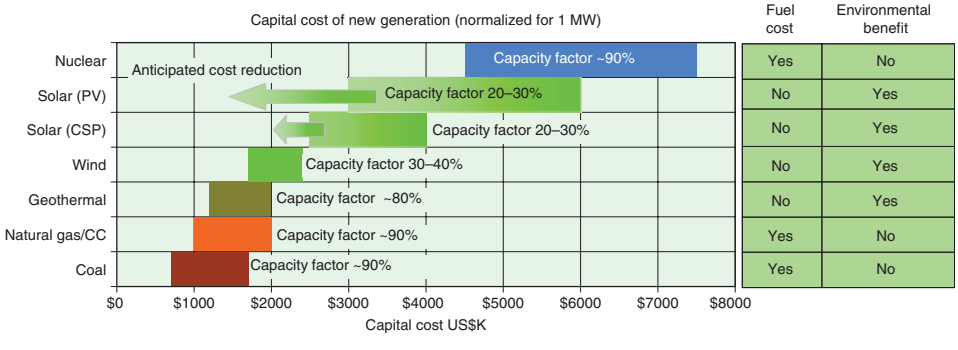


Figure 9.5 Next-generation cost comparison. Source: Reproduced from [540] with permission of ©IEEE.

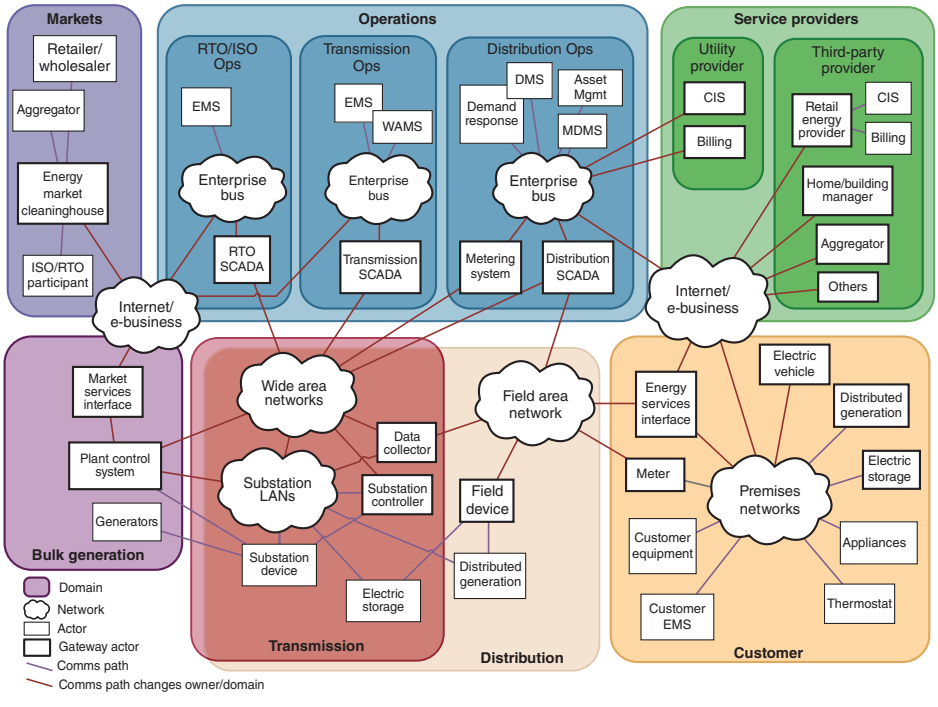


Figure 9.6 NIST smart grid reference model. Source: Reproduced from NIST Report, http://www.nist.gov/smartgrid/upload/NIST-SP-1108r3.pdf. (accessed September 9, 2016).

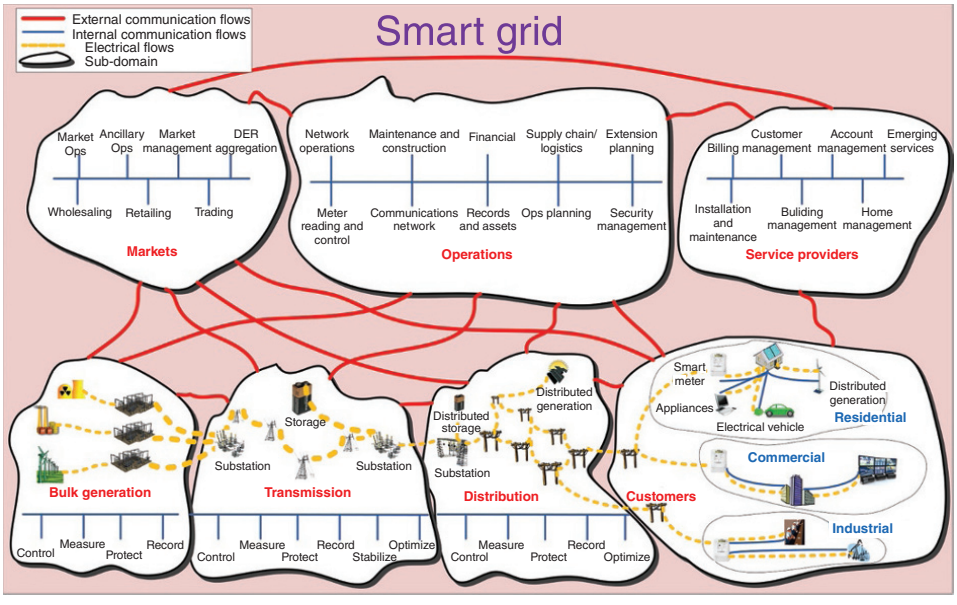


Figure 9.7 Smart grid domains. Source: Reproduced from [533].

Table 9.2 The smart grid compared with the existing grid.

Existing grid	Intelligent grid
Electromechanical	Digital
One-way communication	Two-way communication
Centralized generation	Distributed generation
Hierarchical	Network
Few sensors	Sensors throughout
Blind	Self-monitoring
Manual restoration	Self-healing
Failures and blackouts	Adaptive and islanding
Manual check/test	Remote check/test
Limited control	Pervasive control
Few customer choices	Many customer choices

Source: From [542].

Table 9.3 Evolution of the power system from a static to a dynamic infrastructure.

From	To
Central generation and control	Central and distributed generation with intelligence
Load flow following Kirchoff's law	Load flow control by power electronics
Power generation according to load demand	Controllable generation, variable in feed and demand in equilibrium
Manual switching and trouble response	Automatic response and predictive avoidance
Deterministic response to power flow	Monitored overload of bottlenecks
Periodic maintenance	Prioritized condition-based predictive maintenance

Source: From [541].

- the system is not flexible, so that it is difficult either to inject electricity from alternative sources at any point along the grid, or to efficiently manage new services desired by the users of electricity.

9.4 Future Smart Electrical Energy System

Some of the key requirements of the smart grid include [538] the need for it to allow for:

- the integration of renewable energy resources to address global climate change;
- active customer participation to enable far better energy conservation;
- secure communications;

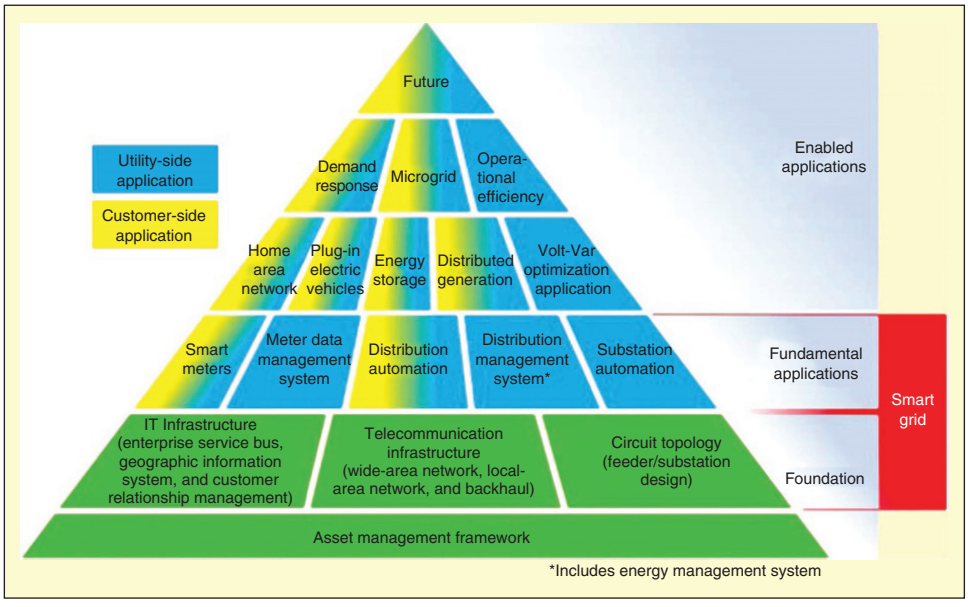


Figure 9.8 Smart grid pyramid. Source: Reproduced from [542] with permission of the IEEE.

- better utilization of existing assets to address long term sustainability;
- optimized energy flow to reduce losses and lower the costs of energy;
- the integration of electric vehicles to reduce dependence on hydrocarbon fuels;
- the management of distributed generation and energy storage to eliminate or defer system expansion and reduce the overall cost of energy;
- allow for the integration of communication and control across the energy system to promote interoperability and open systems and to increase safety and operational flexibility.

The ultimate smart grid is a vision, and it will require cost justification at every step before implementation, then testing and verification before extensive deployment.

The fully implemented smart grid will have the following characteristics [538]:

- Self-healing: automatic repair or removal of potentially faulty equipment from service before it fails, and reconfiguration of the system to reroute supplies of energy to sustain power to all customers.
- Flexible: the rapid and safe interconnection of distributed generation and energy storage at any point on the system at any time.
- Predictive: use of machine learning, weather-impact projections, and stochastic analysis predict the next most likely events so that appropriate actions are taken to reconfigure the system before next worst events can happen.
- Interactive: appropriate information regarding the status of the system is provided not only to the operators, but also to the customers to allow all key participants in the energy system to play an active role in optimal management of contingencies.

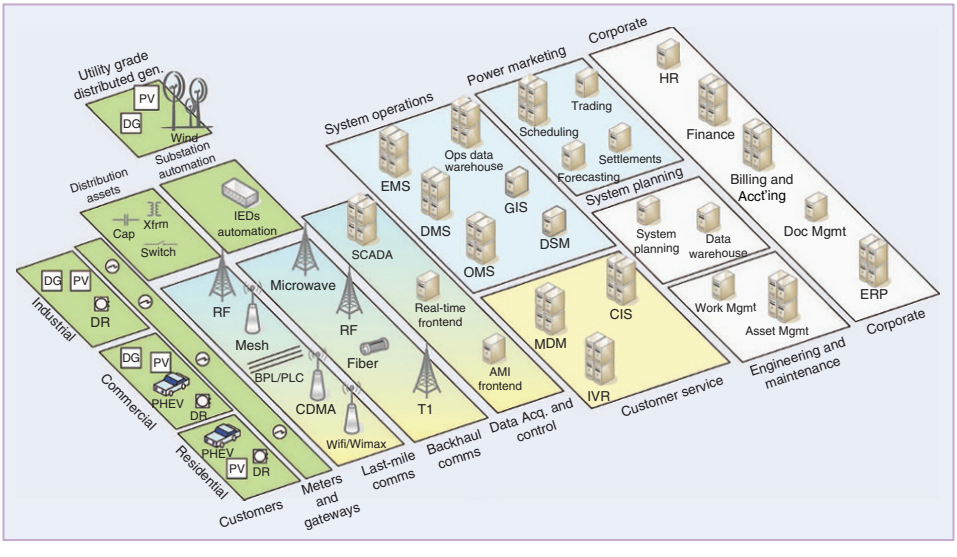


Figure 9.9 A view of the utility information system impacted by smart grid strategies. Source: Reproduced from [540] with permission of the IEEE.

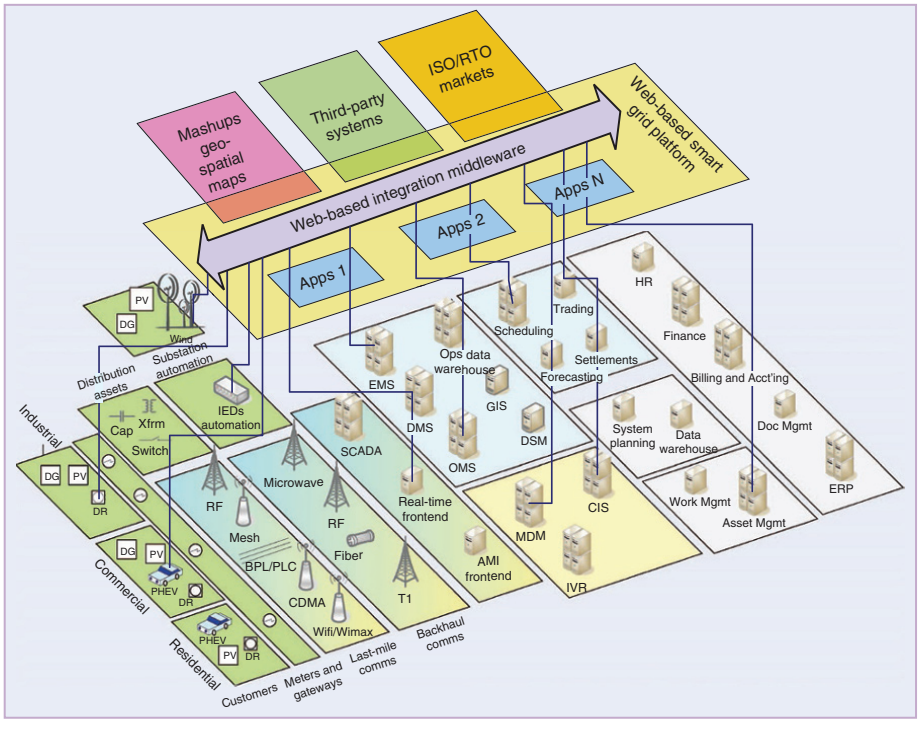


Figure 9.10 Using the cloud for smart-grid applications. Source: Reproduced from [540] with permission of the IEEE.

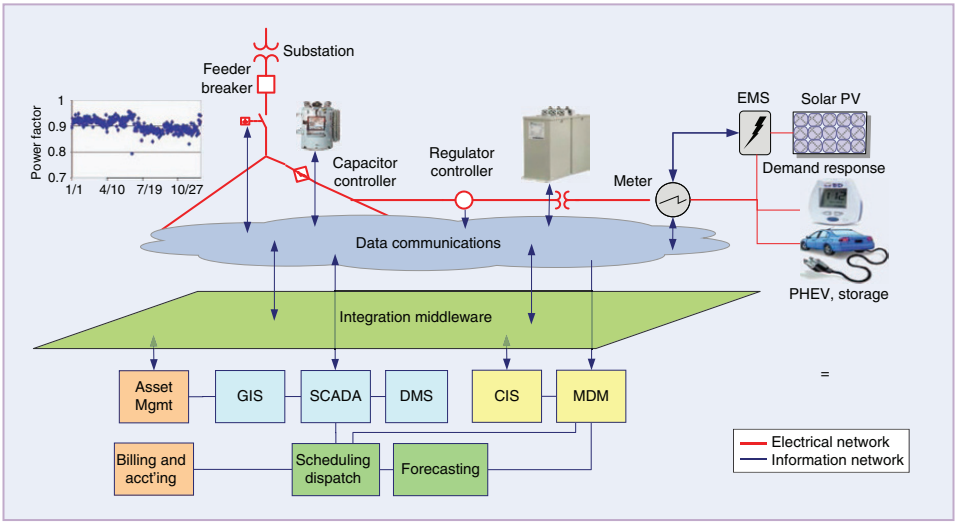


Figure 9.11 Systems required to support the high penetration of distributed resources. Source: Reproduced from [540] with permission of the IEEE.

- Optimized: knowing the status of every major component in real or near real time and having control equipment to provide optional routing paths provide the capability for autonomous optimization of the flow of electricity throughout the system.
- Secure: considering the two-way communication capability of the Smart Grid covering the end-to-end system, there is an essential need for physical and as cyber security of all critical assets.

Table 9.2 and Table 9.3 compare the existing grid with the smart grid. Figure 9.8 gives the smart grid pyramid.

Communications and information technology (IT) are critical to the smart grid. The smart grid will ultimately involve networking vast numbers of sensors in transmission and distribution facilities, smart meters, SCADA systems, back-office systems, and devices in the home, which will interact with the grid. Large amounts of data (hence big data) will be generated by meters, sensors, and synchrophasors [543].

Figure 9.9 shows a view of the utility information system impacted by smart-grid strategies.

One of the emerging and, perhaps, game-changing developments in the IT industry has been the use of the Web (the cloud) as the computing and information management platform. This will allow the integration of data and capabilities from multiple, diverse sources to deliver powerful composite applications over the Web [540]. Figure 9.10 provides a conceptual illustration of this model.

In addition to advanced metering and utilitywide communications infrastructure enabling demand response, and distributed resource management, the smart grid impacts many of the operational and enterprise information systems, including supervisory control and data acquisition (SCADA), feeder and substation automation, customer-service systems, planning, engineering and field operations, grid operations, scheduling, and power marketing. The smart grid also impacts corporate enterprise systems for asset management, billing and accounting, and business management. As illustrated in Figure 9.11, many information technology (IT) systems will be impacted, including those for distribution management and automation, operations planning, scheduling and dispatch, market operations, and billing and settlements.

China prefers to use a "supply-side policy", with a focus on "public enterprise, scientific and technical development and legal regulatory" policies. The United States on the other hand, prefers to use an "environmental-side policy", with a focus on "scientific and technical development, financial, political and public enterprise" policies [544].

10

Technical Challenges for Smart Grid

The potential ramifications of grid failures have never been greater as transport, communications, finance, and other critical infrastructures depend on secure, reliable electricity supplies for energy and control. Several specific pertinent grand challenges to our power systems, economics, and control community persist, including [539]:

- the lack of transmission capability;
- grid operation in a competitive market environment (open access created new and heavy, long-distance power transfer for which the grid was not designed);
- the redefinition of power-system planning and operation in the competitive era;
- the determination of the optimum type, mix, and placement of sensing, communication, and control hardware;
- the coordination of centralized and decentralized control.

10.1 The Conceptual Foundation of a Self-Healing Power System

The Electric Power Research Institute (EPRI)/DoD Complex Interactive Networks/Systems Initiative (CIN/SI) aimed to develop modeling, simulation, analysis, and synthesis tools for the robust, adaptive, and reconfigurable control of the electric power grid and infrastructures connected to it. The intelligent flight control system was designed to provide consistent handling response to the pilot under normal conditions and during unforeseen damage to the aircraft or failure conditions (Figure 10.1).

The damage-adaptive intelligent flight-control system laid the conceptual foundation of a self-healing power system [539], where analogously a squadron of aircraft can be viewed in the same manner as components of a larger interconnected power-delivery infrastructure, a system in which system stability and reliability must be maintained under all conditions, even when one ($N-1$ contingency) or more ($N-k$ contingencies) components are disabled.

The extensions of CIN/SI are discussed in [545].

Smart Grid using Big Data Analytics: A Random Matrix Theory Approach, First Edition.
Robert C. Qiu and Paul Antonik.

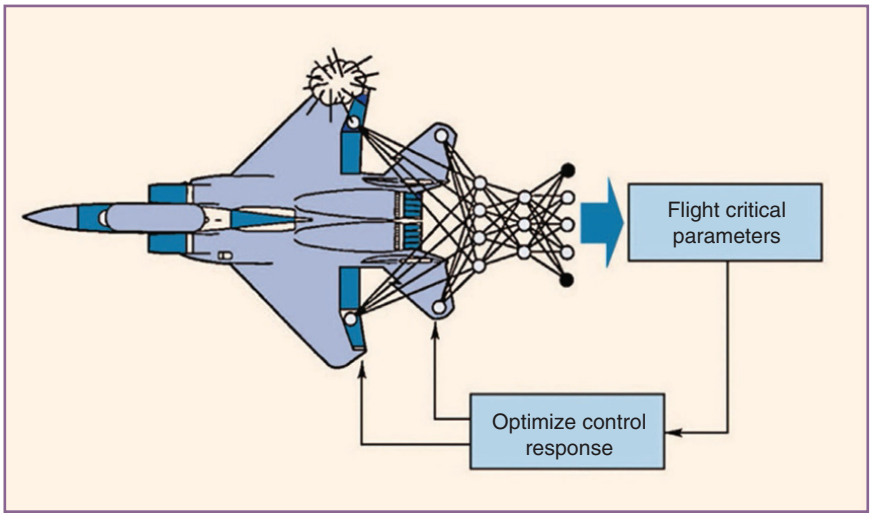


Figure 10.1 A damage-adaptive intelligent flight-control system (IFCS). Source: Reproduced from cite [539] with permission of IEEE.

Table 10.1 A comparison of the protection systems, smart grid, and central control system.

Protection systems	Smart grid	SCADA/EMS central control systems
Local	Fast	SCADA system gathers system status and analog measurements information
Very fast	Distributed	Topology of the power system to determine islands and locate split buses
Few connections to other protection systems	Accurate	Alarms
	Secure	State estimation
	Intelligent	Contingency analysis
		Security dispatch using optimal power flow (OPF)

Source: From [539].

10.2 How to Make an Electric Power Transmission System Smart

Power transmission systems suffer from the fact that intelligence is only applied locally by protection systems and by central control through the supervisory control and data acquisition (SCADA) system. In some cases, the central control system is too slow, and the protection systems (by design) are limited to protection of specific components only [539].

To add intelligence to an electric power transmission system, we need to have independent processors in each component and at each substation and power plant. Table 10.1 compares the smart grid to the protection systems and SCADA/energy management system (EMS) central systems.

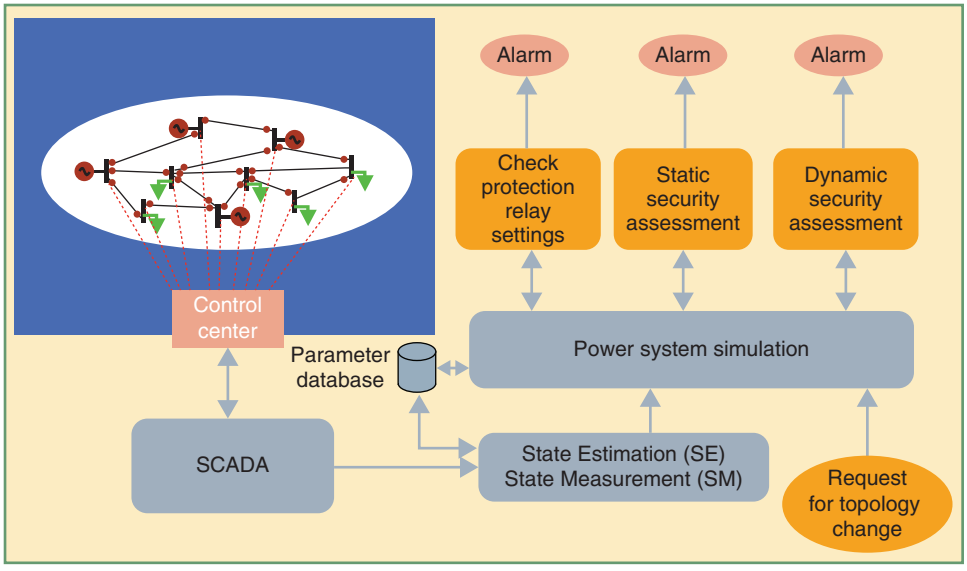


Figure 10.2 How energy management systems can help to avoid blackouts. Source: Reproduced from [541] with permission of the IEEE.

Modern computer and communications technologies now allow us to think beyond the existing protection systems and the central control systems to a fully distributed system that places intelligent devices at each component, substation, and power plant. This distributed system will enable us to build a truly smart grid.

The evolution to smart grids will bring about significant changes to the functionality of energy management and control systems. With the eventual deployment of phasor measurement units to monitor grid performance across heavily loaded regional interconnections, there will be advances in state estimators that are capable of real-time simulations for large networks. As a reuslt, these innovations will assist operators in avoiding major blackouts in the future (Figure 10.2).

10.3 The Electric Power System as a Complex Adaptive System

The electric power grid, made up of many geographically dispersed components, is itself a complex adaptive system that can exhibit global change almost instantaneously as a result of local actions. The Electric Power Research Institute (EPRI) utilized a complex adaptive system to develop modeling, simulation, and analysis tools for adaptive and reconfigurable control of the electric power grid [539]. The underlying concept for the self-healing, distributed control of an electric power system involves treating the individual components as independent intelligent agents, competing and cooperating to achieve global optimization in the context of the whole system environment.

The design includes modeling, computation, sensing, and control. At the system level, each agent in a substation or power plant knows its own state and can communicate with its neighboring agents in other parts of the power system.

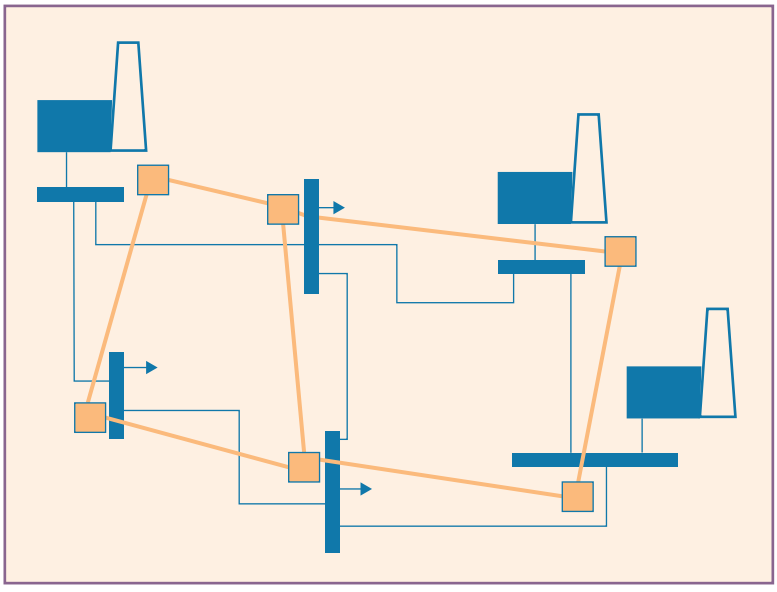


Figure 10.3 A sample system with processors connected by communication links. Source: Reproduced from [539] with permission from the IEEE.

10.4 Making the Power System a Self-Healing Network Using Distributed Computer Agents

In Figure 10.3 we show three power plants connected to load substations through a set of looped transmission lines. Each plant and each substation will have its own processor (designated by a small red box in the figure). Each plant and substation processor is now interconnected in the same manner as the transmission system itself.

How to effectively sense and control a widely dispersed, globally interconnected system is a serious technological problem [539].

10.5 Distribution Grid

In recent decades, more and more stress has been placed on the electricity supply and infrastructure. Electricity usage increased significantly and fluctuated more. Demand peaks have to be generated and transmitted, and they define the minimal requirements in the chain. The goal of our control methodology is to exploit the optimization potential of domestic technologies [546].

A crucial application area for information and communication technology in distribution grids is the fostering of demand-side management (DSM) and demand response programs. It is of vital interest to distribution grid operators to know about the actual grid load and to reshape it if it imperils grid stability. (Figure 10.4). Various price- and incentive-based demand-response programmes have been developed.

Figure 10.5 shows a model of domestic energy streams. Every house consists of (several) microgenerators, heat and electricity buffers, appliances, and a local controller. Multiple houses are combined into a (micro)grid, exchanging electricity and

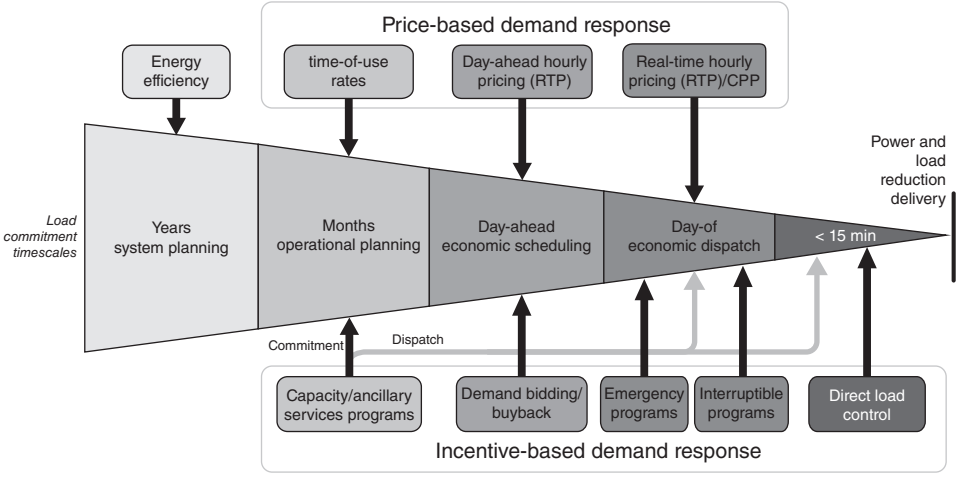


Figure 10.4 Role of demand response in electric system planning and operations. Source: Reproduced from [547].

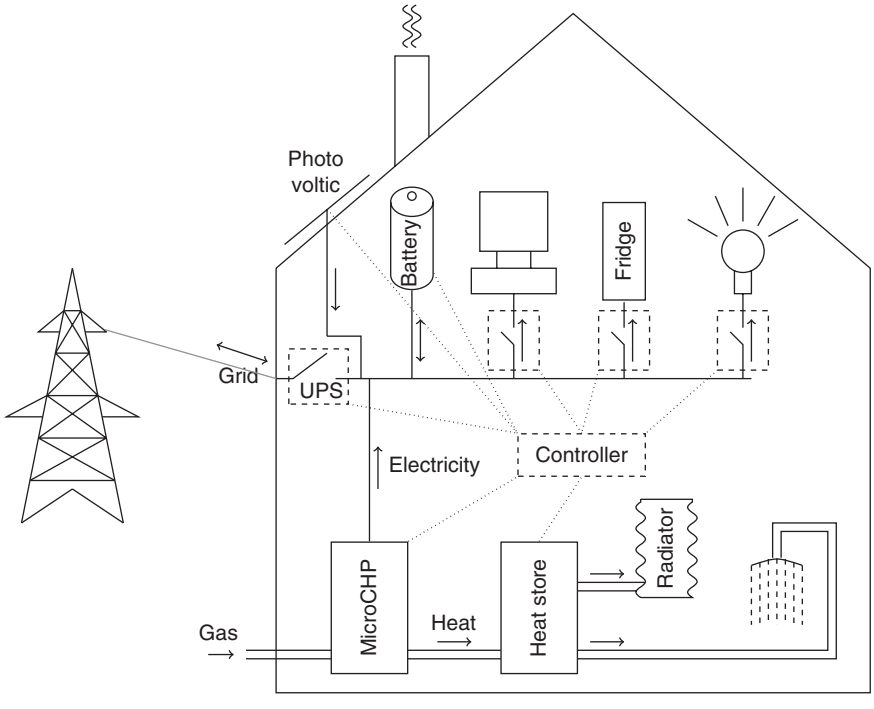


Figure 10.5 Model of domestic energy streams. Source: Reproduced from [546] with permission of IEEE.

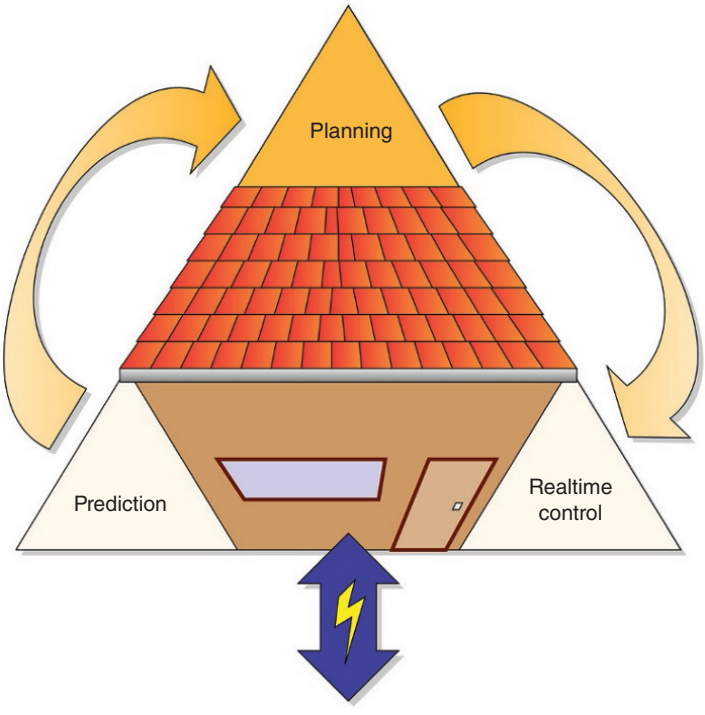


Figure 10.6 Three step control methodology. Source: Reproduced from [546] with permission from IEEE.

information between the houses. Electricity can be imported from and exported into the grid. Heat is produced, stored and used only within the house.

Figure 10.6 shows three-step control methodology. The combination of prediction, planning and real-time control exploits potential. The hierarchical structure with intelligence on the different levels ensures scalability, reduces the amount of communication, and decreases the computation time required for planning.

Since electricity is nonstorable economically, wholesale prices (i.e., the prices set by competing generators to regional electricity retailers) vary from day to day and usually fluctuate by an order of magnitude from low-demand night-time hours to high-demand afternoons. However, in general, almost all retail consumers are currently charged some average price that does not reflect the actual wholesale price at the time of consumption [548].

10.6 Cyber Security

As a critical infrastructure element, smart grid requires the highest levels of security. A comprehensive architecture with security built in from the beginning is necessary. The smart-grid security solution requires a holistic approach including PKI technology elements based on industry standards, and trusted computing elements.

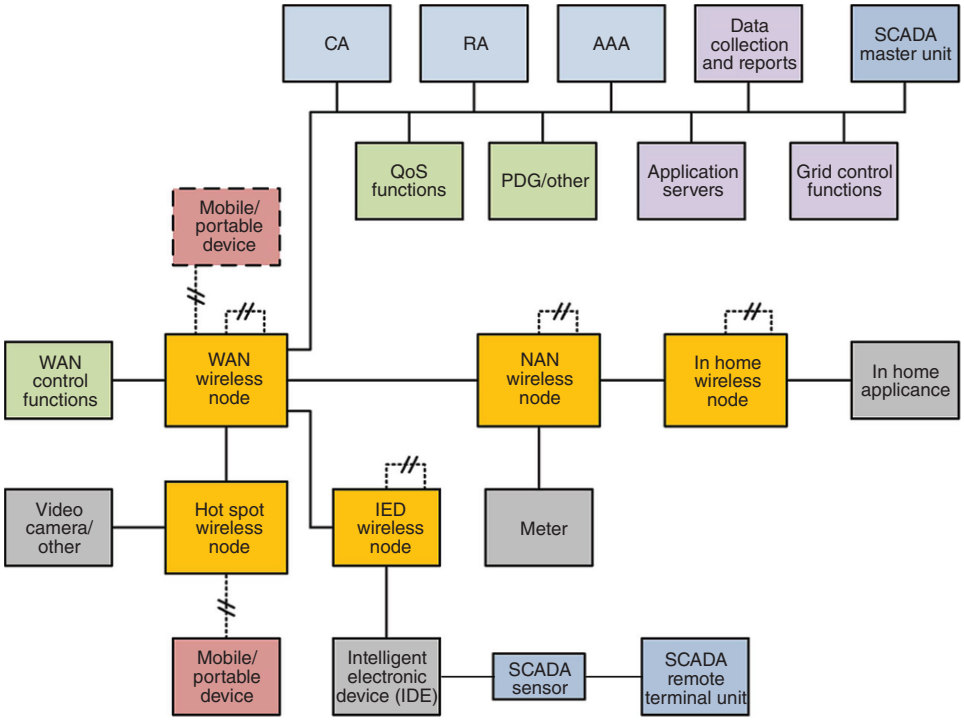


Figure 10.7 Smart grid detailed logical model. Source: Reproduced from [549] with permission from the IEEE.

The diagram in Figure 10.7 shows an example of the possible interconnection of a subset of the various networks, with a WAN wireless network as the backbone of the entire system [549]. Note that the wireless interfaces between similar devices are shown as a dashed, double-hashed line.

10.7 Smart Metering Network

A smart meter is an advanced meter (usually an electrical meter, but it could also integrate or work together with gas, water, and heat meters) that measures energy consumption in much more detail than a conventional meter. Future smart meters will communicate information back to the local utility for monitoring and billing purposes. A smart meter may also potentially communicate with a number of appliances and devices within future smart homes.

Smart meters are expected to provide accurate readings automatically at requested time intervals to the utility company, electricity distribution network or to the wider smart grid. The expected frequency of such readings could be as high as every few (1–5) minutes. Figure 10.8 shows household electricity-demand profile.

Figure 10.9 shows an example for distribution of network smart-metering data.

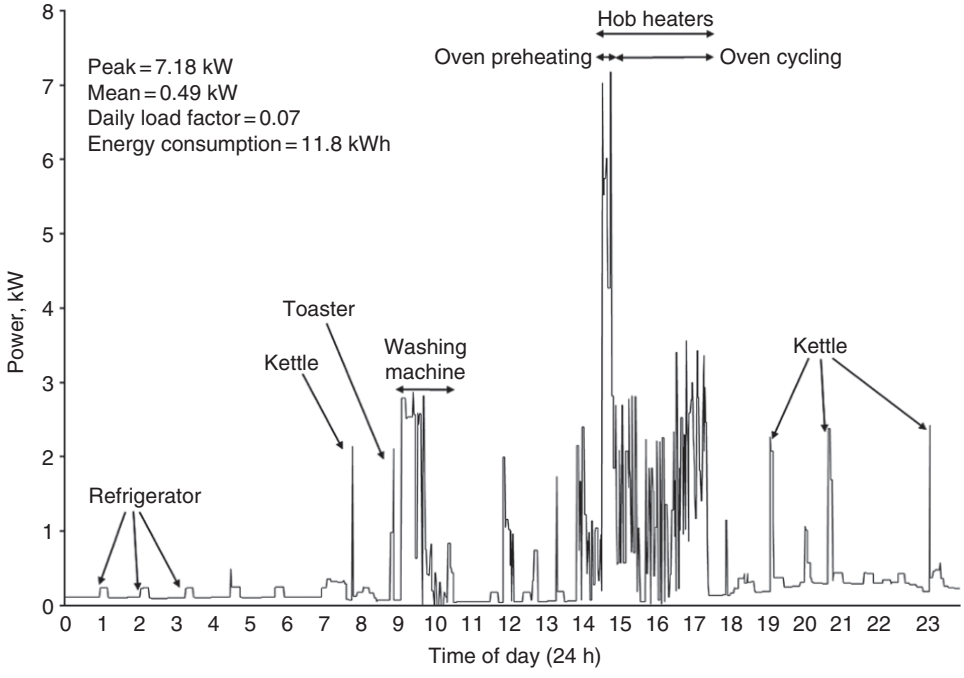


Figure 10.8 Household electricity demand profile. Source: Reproduced from [550] with permission of IEEE.

10.8 Communication Infrastructure for Smart Grid

This section addresses communications topics for smart grids. To control the power electrical grid, we need sensing and communications to tie together the whole grid. High-performance computing and distributed computing are two enablers.

Information plays a crucial role in smart grids, and the communication infrastructure is the decisive component that connects all distributed network elements, enables exchange of information, and therefore makes the grid truly smart [540].

From a communication technology point of view, information exchange in the upper grid levels for SCADA applications is usually covered by existing communication networks belonging to utilities or grid operators (Figure 10.10).

In an increasingly deregulated and distributed energy market, communication between the points of energy generation, distribution, and consumption becomes an essential constituent of efficient grid control [551]. Smart grids rely to a large extent on communication, and the respective infrastructure comprises heterogeneous networks. In order to interconnect them, pure tunneling and gateway approaches are too simple for real-world scenarios. Reference [551] showed that, in reality, a combination of the two plus further tricks are necessary.

- *High reliability and availability.* Nodes should be reachable under all circumstances. It may be challenging for wireless or powerline infrastructures because communication channels can change during operation.

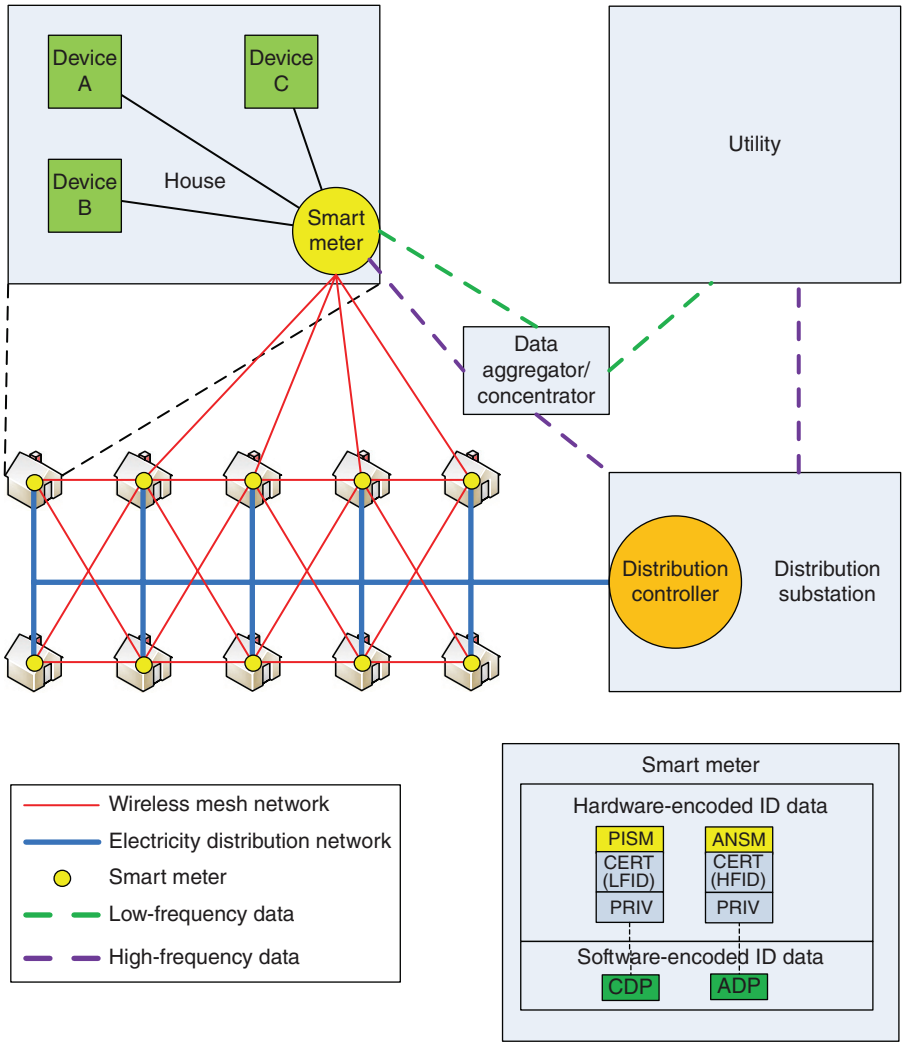


Figure 10.9 Distribution of network smart metering data. Source: Reproduced from [550] with permission of the IEEE.

- *Automatic management of redundancies.* As some applications are time critical, real-time properties of the network have to be maintained even during topology changes.
- *High coverage and distances.* The communication network are distributed in a wide area.
- *Large number of communication nodes.* There are tens of thousands of nodes, particularly in areas of large apartment-block concentration. Even though the commands and data packets are usually short, total data volume to be transferred in the network is substantial, and communication overheads can become an issue.

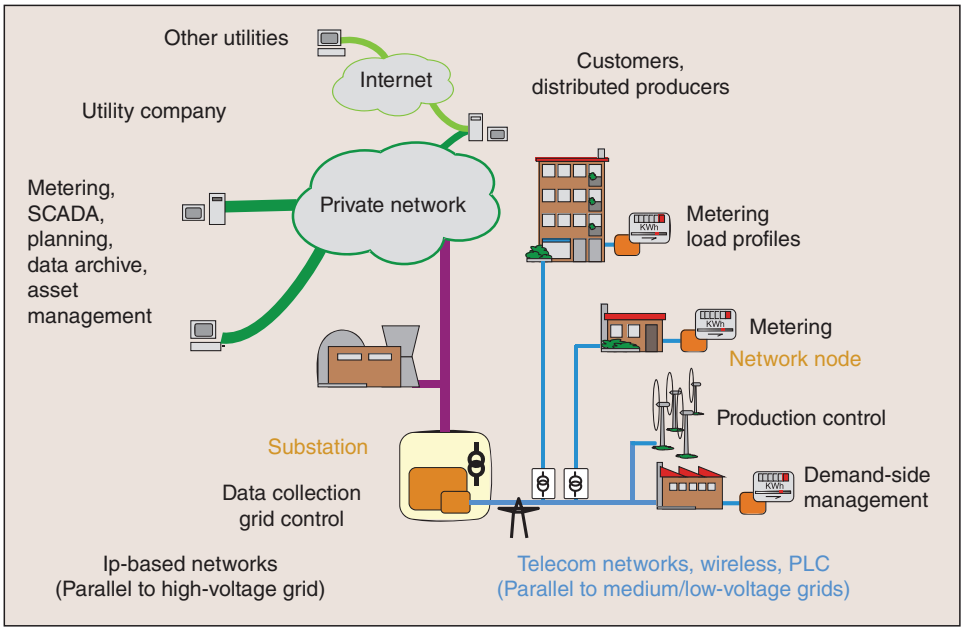


Figure 10.10 Communication Infrastructure for Smart Grid. Source: Reproduced from [535] with permission of the IEEE.

- *Appropriate communication delay and system responsiveness.* The quality-of-service (QoS) management needs to take care of different data classes such as metering, control, or alarm data.
- *Communication security.* Integrity (no malicious modification) and authenticity (origin and access rights are guaranteed) are the most important security goals for energy distribution networks,
- *Ease of deployment and maintenance.*

The existing grid is lack of communication capabilities, while a smart power grid infrastructure is full of enhanced sensing and advanced communication and computing abilities as illustrated in Figure 10.11 and Table 10.2. Different components of the system are linked together with communication paths and sensor nodes to provide interoperability between them, for example distribution, transmission, and other substations, such as residential, commercial, and industrial sites. In the smart grid, reliable and real-time information becomes the key factors for reliable delivery of power from the generating units to the end users.

10.9 Wireless Sensor Networks

The collaborative operation of wireless sensor networks (WSNs) brings significant advantages over traditional communication technologies, including rapid deployment, low cost, flexibility, and aggregated intelligence via parallel processing. The recent advances of WSNs have made it feasible to realize low-cost embedded electric utility

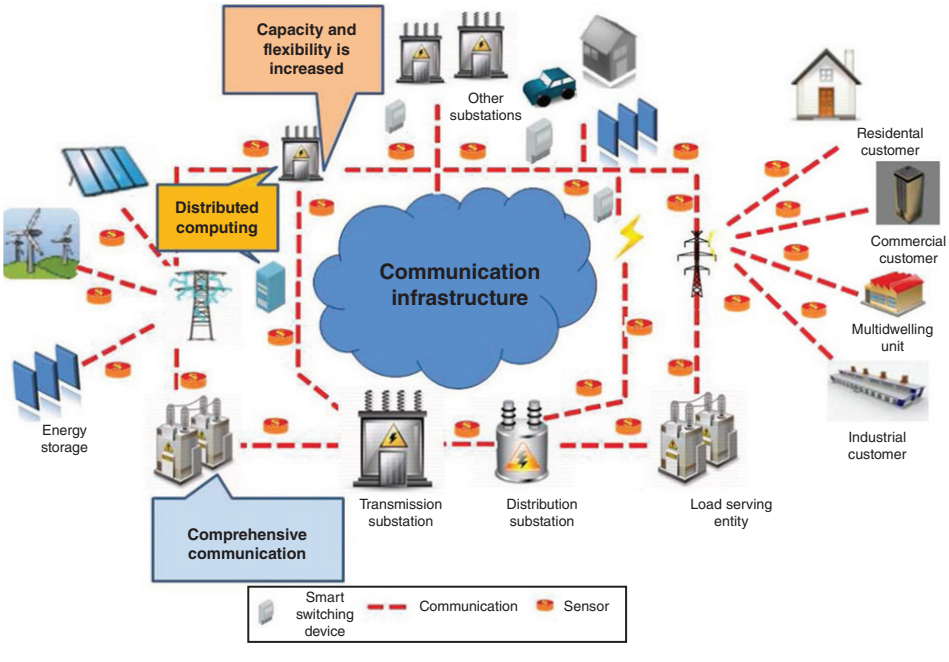


Figure 10.11 Smart grid architecture increases the capacity and flexibility of the network and provides advanced sensing and control through modern communications technologies. Source: Reproduced with permission from [552].

Table 10.2 Smart grid communication technologies.

Technology	Spectrum	Data rate	Coverage range	Applications	Limitations
GSM	900–1800 MHz	Up to 14.4 Kpbs	1–10 1cm	AMI demand response, HAN	Low dale rates
GPRS	900–1800 MHz	Upi to 170 kbps	1–10 km	AMI, demand response, HAN	Low data rates
3G	1.92–1.98 GHz 2.11–2.17 GHz (licensed)	384 Kbps-2 Mbps	1–10 km	AMI, demand response, HAN	Costly spectrum fees
WiMAX	2.5 GHz, 3.5 GHz, 5.8 GHz	Up to 75 Mbps	10–50 km (LOS) 1–5 1cm (NLOS)	AMI, demand response	Not widespread
PLC	1–30 MHz	2–3 Mbps	1–3 km	AMI, fraud detection	Harsh, noisy channel environment
ZigBee	2.4 GHz–868–915 MHz	250 Kbps	30–50 m	AMI, HAN	Low data rate, short range

Source: Reproduced with permission from [552].

monitoring and diagnostic systems. The existing and potential applications of WSNs on smart grid span a wide range, including wireless automatic meter reading (WAMR), remote system monitoring, and equipment fault diagnostics.

Electric power systems contain three major subsystems, power generation, power delivery, and power utilization. Recently, WSNs have been widely recognized as a promising technology that can enhance all these three subsystems, making WSNs a vital component of the next-generation electric power system, the *smart grid*.

The major technical challenges of WSNs in smart-grid applications can be outlined as follows: (i) harsh environmental conditions; (ii) reliability and latency requirements; (iii) packet errors and variable link capacity; (iv) resource constraints.

Bibliographical Remarks

Communications for smart grid is a topic for a whole monograph. See [552] for details. In this book, our approach is to unify many different systems under the umbrella of big data systems.

11

Big Data for Smart Grid

Utilities are on the cusp of a tremendous wave of innovation that will change the way that they operate for ever [553]. We use analytics to prepare for and manage the advent of both the smart grid and the big data age. How do yesterday's AMI and DA deployments affect today's and tomorrow's IT enterprise architectures, command-and-control systems, and next-generation customer services? How will the meteoric growth of electric meter, distributed PV, grid sensor, and electric vehicle data necessitate, influence, or change the software/application layer of smart grid, and the types of systems, platforms, and databases that are relied upon?

11.1 Power in Numbers: Big Data and Grid Infrastructure

Just as intelligent analytics has helped evolve industries, and associated products and services, from IT to healthcare to air travel to social media and online commerce, the same will hold true for the electric power industry, as data is becoming the currency for market transformation.

As the fundamental smart-grid infrastructure continues to be built out via new communication networks and smart hardware, such as meters and control and protection equipment, the stage is being set for an exponential growth in the amount of data that utilities will confront. Challenges, including modeling and simulation, asset management, energy theft detection, DMS/OMS, fault detection and correction, weather data integration, crisis management and mobile workforce management, are revolutionizing grid operations [553].

The majority of smart grid use cases are characterized by the exponential growth of data from the many intelligent communicating devices to be rolled out and the need for fast information retrieval from mass data. The smart grid will be the largest increase in data any energy company has ever seen. The preliminary estimate at one utility is that the smart grid will generate 22 gigabytes of data each day from its 2 million customers [554].

Just collecting the data is not sufficient. Data management has to start at the initial reception of the data, reviewing it for events that should trigger alarms into outage management systems and other real-time systems such as portfolio management of a virtual power-plant operator. The timeframe and volume of available data for information retrieval can reach from real-time data streams to data archived over years [554].

Smart Grid using Big Data Analytics: A Random Matrix Theory Approach, First Edition.
Robert C. Qiu and Paul Antonik.

11.2 Energy's Internet: The Convergence of Big Data and the Cloud

To extract the greatest value from big data, utilities will need the right tools and the right architecture for both their employees and customers so that they can offer self-service (instant Web-based access), speed (in memory analytics) and wide data access and collaboration. Increasingly utilities are turning their attention to the cloud as a way to manage and present new and improved applications, as well as segmenting and prioritizing data. Clouds have already proven very effective in avoiding accidental architecture complexities for utilities looking to pilot new systems without disturbing existing legacy systems. Questions include how, why, when, and what utilities will upload to the cloud, to what degree utilities will rely on both private and public clouds, and how they will continue to rely on and move to cloud-based software-as-a-service products and applications.

11.3 Edge Analytics: Consumers, Electric Vehicles, and Distributed Generation

Consumer engagement is critical for the growth and success of smarter grids. In order to excel with consumer engagement, utilities need to understand consumer behavior through sophisticated analytics [553]. Without a smart grid in place, distributed generation (i.e. renewables), electric vehicles and other consumer advancements would not be able to gain meaningful traction, and to scale to mass penetration and adoption. As solar panels, EVs, and other new grid "assets" begin to "plug in" and communicate, not only will we see the birth of machine-to-machine (M2M) communication; we will also see the immediate need for advanced analytics to manage energy dispatch and usage, voltage irregularities, and other grid operation challenges initiating from the edge of the grid. Challenges lie in the domains of software analytics, grid operations, and renewable energy. The goal is to understand how analytics will *optimize* and *protect* the grid while meanwhile empowering *consumers* to move towards clean energy and Web-based energy management.

11.4 Crosscutting Themes: Big Data

Data collected, analyzed, visualized and warehoused from the smart grid will contribute to many new ideas and inventions that can improve lives [555]. What is needed is a nearly ubiquitous IP transport network operating at bandwidths robust enough to handle traditional utility power delivery applications along with vast amounts of new data from the smart grid. Rather than relying on public communication carriers (AT&T, Sprint, Verizon, etc.), utilities justify the costs of building and operating their own private WANs because of the highly critical nature of these applications for maintaining a reliable and secure power grid.

The enhancement of the monitoring, control, and protection of power systems through smart-grid solutions primarily means availability of more data of better quality than before and availability of new applications that will utilize the data to produce

better decision making [556]. To allow such benefits, the process of data collection, integration, and usage needs to be improved. One example utilizes new data to create improved decision making after occurrence of a fault.

Opportunities and challenges of wireless communication technologies for smart grid applications is treated in [557]. Some potential applications are shown in Figure 11.1.

The Smart Grid will generate billions of data points from thousands of system devices and hundreds of thousands of customers. Data must be converted to useful information through a knowledge-management life cycle in which the data from meters and appliances or substations and distribution systems are analyzed and integrated in a manner that leads to action [555].

The first phase of the knowledge management effort and a key component in the system of information ecology is data conservation in a data warehouse. Data storage needs will explode.

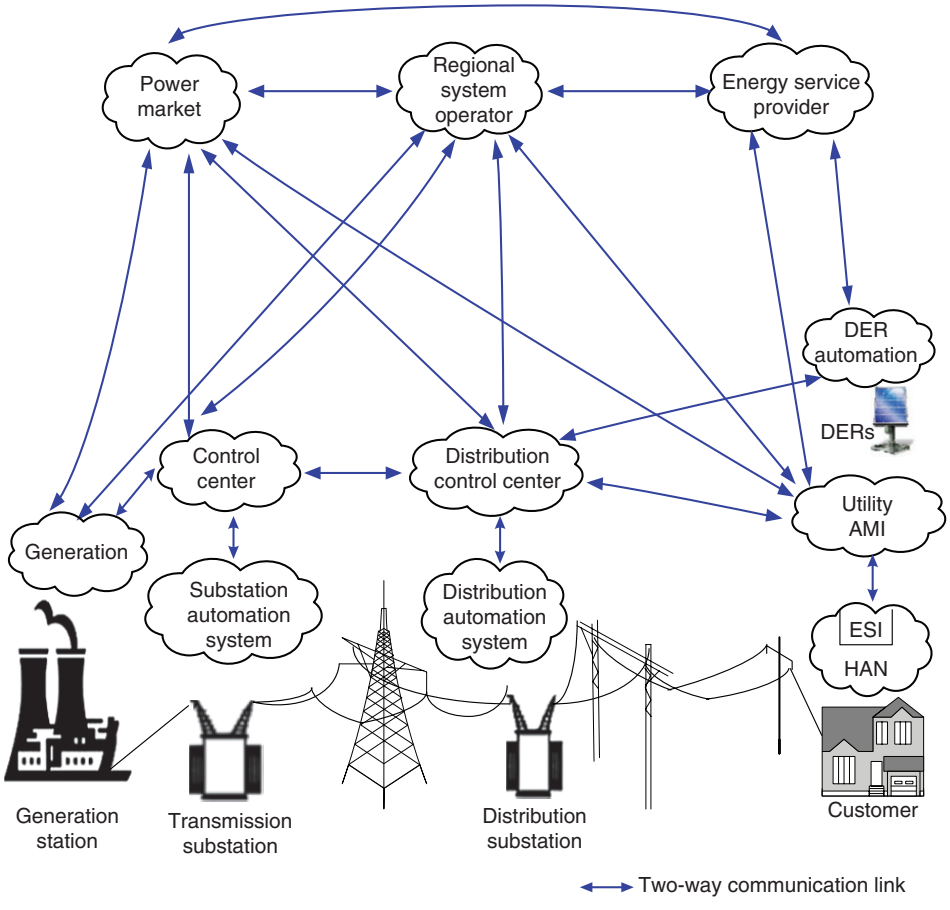


Figure 11.1 Smart grid framework. Source: Reproduced from [557] with permission from IEEE.

11.5 Cloud Computing for Smart Grid

In a cloud-computing environment, flexible data centers offer scalable computing, storage, and network resources to any Internet-enabled device on demand. The drawbacks of traditional information and communication technology in tackling the new challenges are the main obstacle preventing the energy market actors from maximizing the benefits of smart grid as a single entity [554].

Smart-grid operation needs panoramic state data, and, during the operation, maintenance and management of smart grid, massive heterogeneous and multistate data, namely the big data, are generated [558].

11.6 Data Storage, Data Access and Data Analysis

How to store the big data efficiently, reliably and cheaply, and how to access and analyze those data sets rapidly are critical. The source of the big data generated in various process of smart grid, such as power generation, transmission, transformation and power utilization, and the features of the big data are analyzed first.

11.7 The State-of-the-Art Processing Techniques of Big Data

Secondly, existing big-data processing techniques adopted in the fields of business, Internet, and industrial monitoring are summarized, and the advantages and disadvantages of these techniques in coping with the construction of smart grid and big data processing are analyzed in detail.

Finally, aspects of big-data storage, real-time data processing, fusion of heterogeneous multidata sources, visualization of big data, the opportunities and challenges brought by smart-grid big data are expounded [558].

11.8 Big Data Meets the Smart Electrical Grid

Transform the Utility Network

One of the key big data challenges comes with managing data generated by smart meters and smart grids [559]. Issues such as understanding which components of the network are being stressed, determining where future investments can be best made, and identifying which conditions are indicative of future outages can now be addressed.

Pacific Gas & Electric has installed 9 million smart meters to collect data of more than 3 TB. The idea is to lead to innovative uses of information and the ability to pinpoint real-time outages to help the utility more quickly restore power to customers [560]. Big data is a big deal here. The challenges include collecting and effectively analyzing the massive amount of data from smart grid hardware. Another challenge is to correlate the related data, in order to have unprecedented *insight* into how the grid is operating.

Transform Customer Operations

One way to think about smart grid is the convergence of the Internet and a lot of intelligent devices and sensors spread throughout the power system [561]. For example, one of the underlying enabling technologies is advanced metering infrastructure, which puts

smart meters at the end use point. One uses bidirectional communication so the utility can receive information from the meter and communicate to businesses and individual appliances in a home. A single water heater will not make much difference but across 3 million households being able to cut off power to hot water heaters, clothing dryers, and heating and air conditioning adds up.

The objective is to show a high *correlation* between the transaction incentive signal and the variability of wind, how the transaction can interact with heaters and care and analyze the prospects if we scale this up, if instead of 60 000 points of end users there were 3 million.

In addition to basic operational data on transaction incentives and transaction feedback flowing every 5 minutes, we have to collect data from the utilities on the operation of the technology they have bought and installed. Some of the data is related to transactions and some is related to the smart grid operations. All that data flows back to the data center of the operator.

Improve Generation Performance

Generation leads to big data, too [24]. Firstly large data sets are collected and stored in digitized power plant. These data are used for operation analysis, control and optimization, diagnosis analysis, knowledge discovery, and data mining. Secondly data-driven fault diagnosis techniques are used for dynamics systems; using big data for operations, we can obtain new results that are not available when using traditional techniques that are based on models and qualitative empirical knowledge of monitoring. Third, to understand accurately utilities and operation of the distributed generation, we need to monitor and control a large number of distributed sources in real time [562].

By collecting and analyzing key performance and sensor data, it is possible to understand patterns that lead to equipment failure. Big-data analytics supports the change to renewables and microgeneration while providing the necessary mechanism to optimize generation and introduce the necessary demand response capabilities.

Consider one example. In 2006, the DOE and the Federal Energy Regulatory Commission (FERC) recommended that utilities and grid operators install synchrophasor-based transmission-monitoring systems to collect the real-time data needed to predict and manage blackout-related problems, in close to real time, to catch the initial errors and fix them before they lead to disaster. There are up to 6.2 billion data points per day at a size of up to 60 gigabytes with 100 phasor measurement units (PMUs). Increase that to 1000 phasor measurement units, and we obtain up to 41.5 billion data points, or 402 gigabytes of data per day. A lot of that data is flowing into back-end IT systems at the microsecond speed.

11.9 4Vs of Big Data: Volume, Variety, Value and Velocity

The "4Vs" for smart grid are:

1) *Volume.* The volume jumps from TB level to Petabytes (PB) level. In conventional SCADA system, there are 10 000 sampled points. If we collect data every 3–4 s, annually we have a data size of 1.03 TB (1.03 TB = 12 bytes/frame × 0.3 frame/second × 10 000 sampled points × 86,400 seconds/day × 365 days); In a WAMS system, there are 10 000 sampled points, the rate for data collection, however, is increased to 100

times per second (rather than once per 3–4 s); thus, the resultant new data size grows to 495 TB annually.

2) *Variety.* The types of data include: real-time data, historical data, archived data, multimedia data, and time-series data. Some data are structured, semistructured, and unstructured. The requirements for access frequency and data processing speeds are different. The performance requirement is also different.
3) *Value.* In video data, data are collected during continuous monitoring. The useful data may last only for 1–2 s. This is true of the monitoring of transmission utility; the majority of data are normal and a very few of those data are abnormal—the abnormal data are the most important evidence, however.
4) *Velocity.* Within a fraction of second, a massive amount of data be analyzed to support decision making. The performance requirement for online processing is far above the requirement for offline processing. The online analysis and mining of data streams is fundamentally different from the traditional data mining.

11.10 Cloud Computing for Big Data

Cloud computing is part of data storage and processing for big data. Traditional data management is not suitable for big data, which has massive volume and is distributed in nature. The core of cloud computing is data storage for massive data and parallel processing. Google uses distributed file system (DFS) and MapReduce technology (first proposed in 2004).

Designed using low-cost hardware, DFS is highly tolerant of errors and provides high access to data. So the DFS is suitable for computer programs using large data sets. MapReduce is a programming model and an associated implementation for processing and generating large datasets that are amenable to a broad variety of real-world tasks [25]. Users specify the computation in terms of a *map* and a *reduce* function, and the underlying runtime system automatically parallelizes the computation across large-scale clusters of machines, handles machine failures, and schedules intermachine communication to make efficient use of the network and disks. Their work provides a simple and powerful interface that enables automatic parallelization and distribution of large-scale computations, combined with an implementation of this interface that achieves high performance on large clusters of commodity PCs. The programming model can also be used to parallelize computations across multiple cores of the same machine.

Hadoop [563] includes the open-source implementation of MapReduce.

11.11 Big Data for Smart Grid

Attack of big data is a new challenge [564]. The use of big data for smart grid is in its infancy [564]. The progress may be applications driven, rather than technology driven. Visualization is critical to big data in smart grid [565].

A wide-area measurement system (WAMS) is a real-time dynamic power grid monitoring system. Massive WAMS log data processing is based on the platform of Hadoop [566].

11.12 Information Platforms for Smart Grid

The information platform for smart grid based on cloud computing is discussed in [567]

Bibliographical Remarks

For big data in smart grid, see [24].

12

Grid Monitoring and State Estimation

This chapter introduces grid monitoring and state estimation using phasor measurement units (PMUs).

A cornerstone of the smart grid is the advanced monitorability on its assets and operations. Increasingly pervasive installation of PMUs allows so-called synchrophasor measurements to be taken roughly 100 times faster than the legacy supervisory control and data acquisition (SCADA) measurements, time-stamped using the global positioning system (GPS) signals to capture the grid dynamics. On the other hand, the availability of low-latency two-way communication networks will pave the way to high-precision, real-time grid state estimation and detection, remedial actions upon network instability, and accurate risk analysis and postevent assessment for failure prevention.

12.1 Phase Measurement Unit

Synchronized PMUs were first introduced in the early 1980s, and since then have become a mature technology with many applications that are currently under development around the world. Phasor measurement units are power system devices that provide synchronized measurements of real-time phasors of voltages and currents. Synchronization is achieved by same-time sampling of voltage and current waveforms using timing signals from the GPS. The occurrence of major blackouts in many major power systems around the world has given a new impetus for large-scale implementation of wide-area measurement systems (WAMS) using PMUs and phasor data concentrators (PDCs) in a hierarchical structure. Synchronized phasor measurements elevate the standards of power system monitoring, control, and protection to a new level [568].

Data provided by the PMUs are very accurate and enable system analysts to determine the exact sequence of events that have led to the blackouts [569]. One of the most important issues that need to be addressed in the emerging technology of PMUs is site selection [570]. Synchronized phasor measurements have enabled effective and accurate monitoring of the condition of the network in real time with latencies of the order of milliseconds [568].

Smart Grid using Big Data Analytics: A Random Matrix Theory Approach, First Edition.
Robert C. Qiu and Paul Antonik.

12.1.1 Classical Definition of a Phasor

A pure sinusoidal waveform can be represented by a unique complex number known as a phasor. Consider a sinusoidal signal

$$x(t) = X_m \cos(\omega t + \phi) \tag{12.1}$$

The phasor representation of this sinusoid is given by

$$X \equiv \frac{X_m}{\sqrt{2}} e^{j\phi} = \frac{X_m}{\sqrt{2}} (\cos\phi + \sin\phi) \tag{12.2}$$

The signal frequency ω is not explicitly defined in the phasor representation.

12.1.2 Phasor Measurement Concepts

The most common technique for determining the phasor representation of an input signal is to use data samples taken from the waveform, and apply the discrete Fourier transform (DFT) to compute the phasor.

If $x_k \{k = 0, 1, \ldots, N-1\}$ are the N samples of the input signal taken over period, then the phasor representation is given by

$$X = \frac{\sqrt{2}}{N} \sum_{k=0}^{N-1} x_k e^{-jk\frac{2\pi}{N}} \tag{12.3}$$

12.1.3 Synchrophasor Definition and Measurements

In order to obtain simultaneous measurement of phasors across a wide area of the power system, we must synchronize these time tags. As a result, all phasor measurements belonging to the same time tag are truly simultaneous. The PMU must then provide the phasor given by (12.2) using the sampled data of the input signal.

The synchronization is achieved by using a sampling clock which is phase-locked to the one-pulse-per-second signal provided by a GPS receiver. The receiver may be built in the PMU, or may be installed in the substation and the synchronizing pulse distributed to the PMU and to any other device which requires it.

Figure 12.1 shows compensating for signal delay introduced by the antialiasing filter.

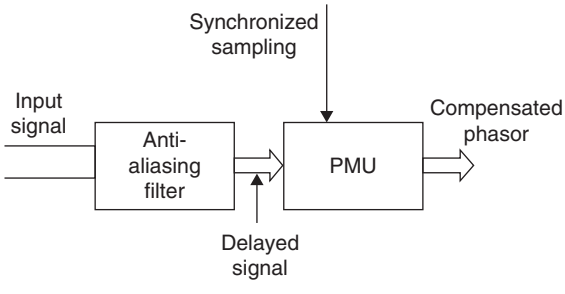


Figure 12.1 Compensating for signal delay introduced by the antialiasing filter. Source: Reproduced from [569] with permission of the IEEE.

12.2 Optimal PMU Placement

When a PMU is placed at a bus, it can measure the voltage phasor at that bus, as well as at the buses at the other end of all the incident lines, using the measured current phasor and the known line parameters [571].

We formulate the problem of determining the minimum number and the optimal locations of the PMUs in terms of an integer quadratic programming approach. The topology of a power system can be expressed by its connectivity matrix $\mathbf{H}$, whose elements are

$$h_{ij} = \begin{cases} 1, & \text{if } i = j \\ 1, & \text{if bus } i \text{ and bus } j \text{ are connected} \\ 0, & \text{otherwise.} \end{cases} \tag{12.4}$$

The binary vector $\mathbf{x} \in \mathbb{R}^n$ of PMU placements is defined as

$$x_i = \begin{cases} 1, & \text{if a PMU is placed at bus } i \\ 0, & \text{otherwise} \end{cases} \tag{12.5}$$

The entries of the product $\mathbf{Hx}$ represent the number of times a bus is observed by the PMU placement set defined by $\mathbf{x}$. The objective function V (x) for optimization is formulated as in an integer quadratic programming problem

$$J(\mathbf{x}) = \gamma (\mathbf{N} - \mathbf{Hx})^T \mathbf{R} (\mathbf{N} - \mathbf{Hx}) + \mathbf{x}^T \mathbf{Q} \mathbf{x} \tag{12.6}$$

where $\gamma \in \mathbb{R}$ is a weight, and $\mathbf{N} \in \mathbb{R}^n$ is a vector representing the upper limits of the number of times each bus can be observed. The diagonal matrix $\mathbf{R} \in \mathbb{R}^{n \times n}$ has entries r_{ii} representing the "significance" of each bus i, and the diagonal matrix $\mathbf{Q} \in \mathbb{R}^{n \times n}$ has entries q_{ii} representing the cost of placing a PMU at bus i. In the generic case where all buses are equally significant and the PMU installation costs at each bus are the same, $\mathbf{Q}$ and $\mathbf{R}$ are equal to the identity matrix $\mathbf{I}^{n \times n}$.

12.3 State Estimation

State estimators (SE) constitute the cornerstone of modern energy management systems, where diverse applications rely on accurate information about the system state [538].

12.4 Basics of State Estimation

The crystallizing vision of the smart grid aspires to build a cyber-physical network that can address these challenges by capitalizing on state-of-the-art information technologies in sensing, control, communication, optimization, and machine learning [573]. Advanced metering systems are needed; also data communication networks throughout the grid are needed. As a result, algorithms that optimally exploit the pervasive sensing and control capabilities of the envisioned advanced metering infrastructure (AMI) are needed to make the necessary breakthroughs in the key problems in power grid monitoring and energy management.

Since the pioneering work of F. C. Schweppe in 1970 [574], state estimation has become a key function in supervisory control and planning of electric power grids [575]. It serves to monitor the state of the grid and enables energy-management systems (EMS) to perform various important control and planning tasks such as establishing near real-time network models for the grid, optimizing power flows, and bad data detection/analysis (see, for example [576] and [577] and the references therein). Another example of the utility of state estimation is the state estimation-based reliability/security assessment deployed to analyze contingencies and determine necessary corrective actions against possible failures in the power systems.

There are at least three major aspects in the future power grid that will directly affect state estimation research. First, more advanced measurement technologies like phase measurement units (PMUs) have offered hope for near real-time monitoring of the power grid; see [578]. Typically, a PMU takes 30 measurements per second, thereby giving a much more timely view of the power system dynamics than conventional measurements. More importantly, all PMUs' measurements are synchronized; they are time stamped by the GPS's universal clock. The PMUs' higher measurement frequency, however, put pressure on the communication and data-processing infrastructure of the grid.

Second, new regulations and market pricing competition may require utility companies to share information and monitor the grid over large geographical areas. This calls for distributed control and thus distributed state estimation to facilitate interconnect-wide coordinated monitoring [579].

Finally, to facilitate smart grid features such as demand response and two-way power flow, utility companies will need to have more timely and accurate models for their distribution systems.

12.5 Evolution of State Estimation

Figure 12.2 gives the electricity ecosystem of the future grid featuring various players and levels of interaction [575]. We envision that state estimation in the future grid will likely be performed at different levels; specifically, the transmission system operator (TSO) level, the local level or substransmission level, and the distribution level; see, for example the multilevel state estimation paradigm [580]. The TSO is an entity that operates the transmission grid to supply electricity from the generating companies (GENCOs) [581] to the utility companies and then to the consumer. Substations are a vital link between the transmission and distribution networks and are responsible for converting voltage and current levels. The trend of deregulation of vertically integrated utilities, particularly in the United States, would mean that market forces would play an increasing role in the future grid. The state of a power system can be described by the voltage magnitudes and phase angles at every bus. This information, along with the knowledge of the topology and impedance parameters of the grid, can be used to characterize the entire system. The EMS/supervisory control and data acquisition (SCADA) system is a set of computational tools used to monitor, control, and optimize the performance of a power system. State estimation is a critical component here; the relationship between state estimation and the SCADA system is shown in Figure 12.3. The data acquisition system obtains measurements from devices like remote terminal units (RTUs)

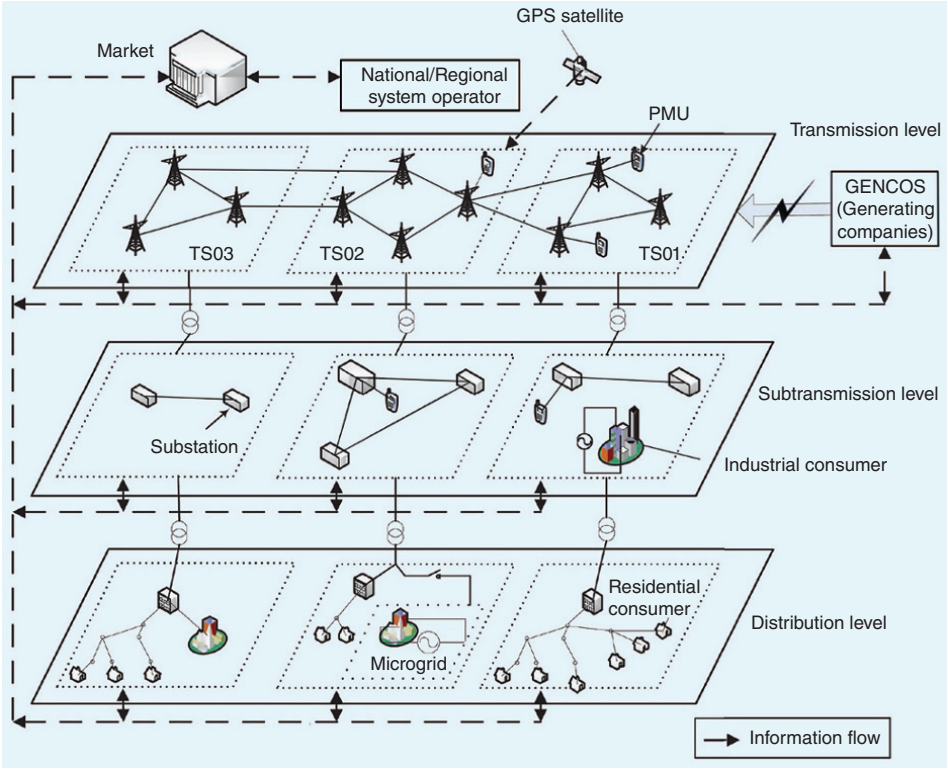


Figure 12.2 Electricity ecosystem of the future grid featuring various of players and levels of interactions. Reproduced from [575] with permission of the IEEE.

and, more recently, phasor data concentrators (PDCs). The state estimator calculates the system state and provides the necessary information to the supervisory control system, which then takes action by sending control signals to the switchgear (circuit breakers).

Depending on the timing and evolution of the estimates, state estimation schemes may be classified into two basic distinct paradigms: static state estimation and forecasting-aided state estimation.

12.6 Static State Estimation

Since the mid-1970s, much of the research on state estimation has been focused on static state estimation, primarily due to the fact that the traditional monitoring technologies, such as those implemented in the SCADA system, can only take nonsynchronized measurements once every 2 to 4 s. To reduce the computational complexity required in implementing state estimation, the estimates are usually updated only once every few minutes. As a result, the usefulness of static state estimation as a means to provide real-time monitoring of the power grid is quite limited in practice.

State estimation [582] processes the whole set of measurements globally and takes advantage of its redundancy to detect any data errors. In this section we first review

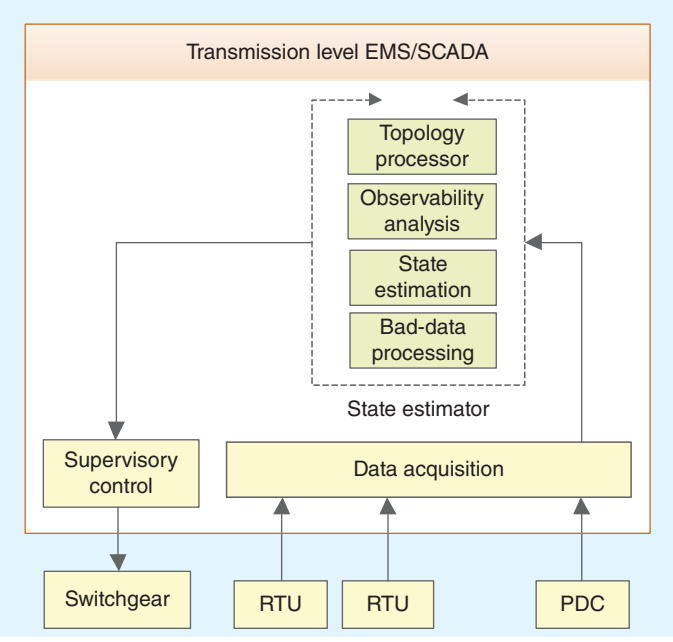


Figure 12.3 Relationship between different elements that collectively constitute the EMS/SCADA. Source: Reproduced from [575] with permission of the IEEE.

the classical formulation of the state estimation using the normal equations, and the linearized decoupled version of it. The nonlinear equations relating the measurements z and the state vector x are

$$\mathbf{z} = \mathbf{h}(\mathbf{x}) + \mathbf{e} \tag{12.7}$$

where $\mathbf{e}$ is the measurement error vector, which is assumed to be jointly Gaussian.

In an N-bus system, the $(2N-1)\times 1$ state vector has the form $\mathbf{x} = \left[\theta_2, \theta_3, \ldots, \theta_N, \left|V_1\right|, \ldots, \left|V_N\right|\right]^T$ where θ_i denotes the phase angles and $\left|V_i\right|$ denotes the magnitudes of the voltages at the i-th bus. The phase angle θ_1 at the reference bus is assumed to be known and is normally set to zero radians. To estimate the state $\mathbf{x}$, a set of measurements $\mathbf{z} \in \mathbb{R}^{L\times 1}$, $L > 2N-1$, is collected. These measurements consist of nonsynchronized active and reactive power flows in network elements, bus injections, and voltage magnitudes at the buses. The measurements are typically obtained within SCADA systems, and related to the state vector by an overdetermined system of nonlinear equations, (12.8).

The state estimation problem is formulated mathematically as a weighted least-square problem and solved by an iterative scheme [582]. At each iteration the procedure is equivalent to solving a linearized weighted least-square (WLS) problem. We may further decouple the real and reactive part of the measurements and the state vector [583, 584]. The resulting linearized decoupled state estimator solves two linear weighted least-square problems of the following form:

$$\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{e} \tag{12.8}$$

where, for the real power case, $\mathbf{z}$ is the set of real power measurements, $\mathbf{x}$ is the set of real part of the state vector (bus angles), $\mathbf{H}$ is the Jacobian matrix of the real measurements with respect to phase angles. Similarly for the reactive case. The solution of the WLS problem (12.8) is

$$(\mathbf{H}^T\mathbf{W}\mathbf{H})\hat{\mathbf{x}} = \mathbf{H}^T\mathbf{W}\mathbf{z} \tag{12.9}$$

or

$$\hat{\mathbf{x}} = (\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{W}\mathbf{z} \tag{12.10}$$

where $\hat{\mathbf{x}}$ is the estimated state, $\mathbf{W}$ is the diagonal matrix of weighting factors (i.e. inverse of the covariance matrix of $\mathbf{e}$), and $(\mathbf{H}^T\mathbf{W}\mathbf{H})$ is called the gain matrix.

The residual vector r is defined to be the difference between the measured quality and the calculated value from the estimated state:

$$\mathbf{r} = \mathbf{z} - \mathbf{H}\hat{\mathbf{x}} \tag{12.11}$$

After some manipulation, we easily obtain

$$\mathbf{r} = (\mathbf{I} - \mathbf{M})\mathbf{e} \tag{12.12}$$

$$\mathbf{M} = \mathbf{H}(\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{W} \tag{12.13}$$

and the expected value and the covariance of the residual vector $\mathbf{r}$ are:

$$\begin{aligned} &\mathbb{E}\{\mathbf{r}\} = 0 \\ &\mathbb{E}\left\{\mathbf{r}\mathbf{r}^T\right\} = (\mathbf{I} - \mathbf{M})\mathbf{W}^{-1} \end{aligned} \tag{12.14}$$

Topology error can be detected. We assume that errors in the status data of breakers and switches will result in erroneous assertion of network topology in terms of either (i) branch outage, (ii) bus split, or (iii) shunt capacitor/reactor switching.

Example 12.6.1 (topology error) The branch outage includes transmission line or transformer outage. In most practical cases, errors in recognizing line or transformer outages may involve only a single outage. Without topology error, we have $\mathbb{E}(\mathbf{r}) = 0$. If a topology error is present, $\mathbb{E}(\mathbf{r})$ is equal to something else.

The effect of topology error appears in the matrix $\mathbf{H}$. Let $\mathbf{H}$ be the true measurement Jacobian matrix, and $\tilde{\mathbf{H}}$ be the one from the topology processor with errors, and $\Delta\mathbf{H}$ be the resulting error in the measurement Jacobian matrix:

$$\mathbf{H} = \tilde{\mathbf{H}} + \Delta\mathbf{H} \tag{12.15}$$

The true equation for the state estimation should be

$$\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{e} \tag{12.16}$$

However, due to topology error, the following equation instead is obtained for estimating the state:

$$\mathbf{z} = \tilde{\mathbf{H}}\mathbf{x} + \mathbf{e} \tag{12.17}$$

The estimated error $\tilde{\mathbf{x}}$ is

$$\hat{\mathbf{x}} = \left(\tilde{\mathbf{H}}^T\mathbf{W}\tilde{\mathbf{H}}\right)^{-1}\tilde{\mathbf{H}}^T\mathbf{W}\mathbf{z} \tag{12.18}$$

The residual vector can be obtained by substituting (12.21) and (12.18) into (12.19)

$$\mathbf{r} = \mathbf{z} - \tilde{\mathbf{H}}\hat{\mathbf{x}} \tag{12.19}$$

and we have

$$\mathbf{r} = (\mathbf{I} - \mathbf{M})(\Delta\mathbf{H}\mathbf{x} + \mathbf{e}) \tag{12.20}$$

Thus, we have

$$\begin{aligned} \mathbb{E}(\mathbf{r}) &= (\mathbf{I} - \mathbf{M})\,\Delta\mathbf{H}\mathbf{x} \\ \mathrm{Cov}(\mathbf{r}) &= (\mathbf{I} - \mathbf{M})\,\mathbf{W}^{-1} \end{aligned} \tag{12.21}$$

where $\mathrm{Cov}(\mathbf{r})$ is the covariance matrix of the vector $\mathbf{r}$. □

State estimation is based on the hypothesis that there is no gross error ($\mathbf{e}$ is Gaussian) in the measurements. This hypothesis may be tested using the normalized residuals. This, so called $\mathbf{r}^N$ test, is based on the fact that there is no gross error: $\mathbb{E}(\mathbf{r}) = 0$. So, the hypothesis is accepted if

$$\max_i \left| r_i^N \right| < \gamma$$

where γ is the detection threshold. The $\mathbf{r}^N$ test has been shown to be effective in detecting and identifying bad data.

As $\mathbf{e}$ is standardized Gaussian, $\mathbf{r}$ is also Gaussian with covariance $\mathrm{Cov}(\mathbf{r})$; hence, $\|\mathbf{r}\|_2^2$ is a χ^2 distribution with degrees of freedom $(L - 2N + 1)$. The χ^2 test then declares a least square-based power system state estimation possibly affected by outliers whenever $\|\mathbf{r}\|_2^2$ exceeds a predefined threshold.

Recently, a semidefinite relaxation (SDR) approach has been recognized to develop polynomial-time PSSE algorithms with the potential to find a globally optimal solution [585, 586].

12.7 Forecasting-Aided State Estimation

Conventional statistic state estimation relies on a single set of measurements all taken at one snapshot in time. So, it disregards the evolution of the state over consecutive measurement instants. The basic idea of forecasting-aided state estimation is to provide a *recursive update* of the state estimate that can also track the changes occurring during normal system operation. One of the advantages of forecasting-aided state estimation is that it includes by design a forecasting feature that can avoid the problem of *missing measurements*, as the predicted states may be used to replace those missing measurements. Note that forecasting-aided state estimation is somewhat different from the true dynamic state estimation because the transients in power systems are usually in a much faster time scale than those considered in forecasting-aided state estimation. An extensive survey is given in [587].

A typical forecasting-aided state estimation is formulated with the following dynamic model [588]:

$$\mathbf{x}(k+1) = \mathbf{F}(k)\mathbf{x}(k) + \mathbf{g}(k) + \mathbf{w}(k) \tag{12.22}$$

where for time constant k, $\mathbf{F}(k) \in \mathbb{R}^{(2N-1)\times(2N-1)}$ is the state transition matrix, vector $\mathbf{g}(k)$ is associated with the trend behavior of the state-trajectory, and $\mathbf{w}(k)$ is assumed to be zero-mean Gaussian noise with covariance matrix $\mathbf{C}_w$ defined in (12.21); hence $\|\mathbf{r}\|_2^2$ follows a χ^2 distribution with $(m-n)$ degrees of freedom.

Using (12.22) and the measurements arriving at instant $k+1$

$$\mathbf{z}(k+1) = \mathbf{h}(\mathbf{x}(k)) + \mathbf{n}(k+1)$$

where $\mathbf{n}(k+1)$ is a zero-mean Gaussian measurement noise vector with covariance matrix $\mathbf{C}_n \in \mathbb{R}^{L\times L}$, the majority of the forecasting-aided state estimation algorithms that appear in the literature are based on the extended Kalman filter (EFL), whose recursions are given by

$$\hat{\mathbf{x}}(k+1) = \tilde{\mathbf{x}}(k+1) + \mathbf{K}(k+1)\left[\mathbf{z}(k+1) - \mathbf{h}(\tilde{\mathbf{x}}(k+1))\right] \tag{12.23}$$

where

$$\begin{aligned}
\tilde{\mathbf{x}}(k+1) &= \mathbf{F}(k)\hat{\mathbf{x}}(k) + \mathbf{g}(k)\\
\mathbf{K}(k+1) &= \mathbf{\Sigma}(k+1)\mathbf{H}^T(k+1)\mathbf{C}_n^{-1}\\
\mathbf{\Sigma}(k+1) &= \left[\mathbf{H}^T(k+1)\mathbf{C}_n^{-1}\mathbf{H}(k+1) + \mathbf{M}^{-1}(k+1)\right]^{-1}\\
\mathbf{M}(k+1) &= \mathbf{F}(k)\mathbf{\Sigma}(k)\mathbf{F}^T(k) + \mathbf{C}_w
\end{aligned}$$

with $\mathbf{H}(k+1)$ being the measurement Jacobian evaluated at $\tilde{\mathbf{x}}(k+1)$.

Since the power grid is inevitably a large network, a centralized solution to the associated state estimation problem poses tremendous computational complexity. An alternative is to divide the large power system into smaller areas, each equipped with a local processor to provide a local state estimation solution. As compared with a centralized state estimation approach, multiarea state estimation reduces the amount of data that each state estimator needs to process (and hence reduces complexity) and it improves the robustness of the system by distributing the knowledge of the state. However, its implementation requires additional communication overhead and it comes with the time-skewness problem that results from asynchronous measurements obtained in different areas.

Each area has local measurements formulated by

$$\mathbf{z}_m = \mathbf{h}_m\left(\mathbf{x}_m\right) + \mathbf{n}_m, \quad m = 1, \ldots, M \tag{12.24}$$

where $\mathbf{x}_m = \left[\mathbf{x}_{im}^T \mathbf{x}_{bm}^T\right]^T$ is the local state vector of area m, which is further partitioned into internal state variables, $\mathbf{x}_{im}^T$, and border state variables, $\mathbf{x}_{bm}^T$. Internal variables are those state variables that are observable for the particular area while border variables are states of those buses with lines connecting two areas (so-called tie lines).

12.8 Phasor Measurement Units

The PMUs sample at a much higher frequency (roughly *two orders* of magnitude faster) compared to the traditional sensors in the SCADA system. The PMUs provide more accurate and more timely measurements with many more samples. The main challenges faced by engineers today include (i) combining those PMU measurements with conventional measurements to obtain an optimal state estimate, and (ii) dealing with the large

number of data rendered by PMUs. Novel techniques need to be developed to extract relevant state information from the tidal wave of measurement data.

12.9 Distributed System State Estimation

Research on distribution system state estimation dates back to the early 1990s; see, for example [589]. This scheme has not been truly brought into fruition, probably due to the lack of proper infrastructure.

The most popular method used in traditional power system state estimation is the maximum likelihood estimation (MLE). It assumes the state of the system is a set of deterministic variables and determines the most likely state via error included interval measurements. In the distribution system, the measurements are often too sparse to fulfill the system observability. Instead of introducing pseudo-measurements, authors in [590] propose a belief propagation (BP) based distribution system state estimator. This new approach assumes that the system state is a set of stochastic variables. With a set of prior distributions, it calculates the posterior distributions of the state variables via real-time sparse measurements from both traditional measurements and the high resolution smart metering data.

12.10 Event-Triggered Approaches to State Estimation

The future grid will be equipped with a myriad of smart meters, which will collect and transmit massive amount of data, and the control center will need to process those data, convert data into information and transform information into actionable intelligence. In fact, the deployment of PMUs at the transmission level has already resulted in more data than the legacy grid's control center can handle. When the smart grid is fully deployed, the so-called big data phenomenon occurs naturally.

It is desirable to make the communication infrastructure throughout the grid energy and bandwidth efficient. As a result, an event-triggered approach to sensing, communicating and information processing would be quite appealing.

12.11 Bad Data Detection

One of the essential benefits of using a state estimator is the ability to detect, identify, and correct measurement errors. This procedure is referred to as bad data processing. This is especially relevant in the context of cyber security. Depending upon the state estimation method used, bad data processing may be carried out as part of the state estimation or as a postestimation procedure. Irrespective of the state estimation method employed, detectability of bad data depends upon the measurement configuration and redundancy [591].

The purpose of a static state estimator is to find the estimate $\hat{\mathbf{x}}$ of the true state $\mathbf{x}$ that best fits the measurement $\mathbf{z}$ related to $\mathbf{x}$ through the nonlinear model:

$$\mathbf{z} = \mathbf{h}(\mathbf{x}) + \mathbf{w} \tag{12.25}$$

where we have used the customary notation:

- $\mathbf{z}$: the m-dimensional measurement vector;
- $\mathbf{x}$: the n-dimensional state vector of voltage magnitudes and phase angles;
- $n = 2N - 1$, N being the number of system nodes; in estimation, $n < m$, i.e., the redundancy $\eta = m/n > 1$;
- $\mathbf{w}$: the m-dimensional measurement error vector; its i-th component is: (i) a white Gaussian noise $\mathcal{N}\left(0, \sigma_i^2\right)$ if the corresponding measurement is valid; (ii) an unknown deterministic quantity otherwise.

The weighted least square (WLS) estimate $\hat{\mathbf{x}}$ based on the quadratic criterion $J(\mathbf{x})$ satisfies the optimality condition

$$\mathbf{H}^T(\hat{\mathbf{x}})\,\mathbf{R}^{-1}\left(\mathbf{z} - \mathbf{h}(\hat{\mathbf{x}})\right) = \mathbf{H}^T(\hat{\mathbf{x}})\,\mathbf{R}^{-1}\mathbf{r} = 0 \tag{12.26}$$

where $\mathbf{H} \triangleq \partial\mathbf{h}/\partial\mathbf{x}$ represents the Jacobian matrix, $\mathbf{R} = \text{diag}\left(\sigma_i^2\right)$ and the measurement residual vector $\mathbf{r}$ is defined as

$$\mathbf{r} \triangleq \mathbf{z} - \mathbf{h}(\hat{\mathbf{x}}) = \mathbf{Q}\mathbf{w} \tag{12.27}$$

In the latter expression, the residual sensitivity matrix $\mathbf{Q}$ is given by

$$\mathbf{Q} = \mathbf{I} - \mathbf{H}\boldsymbol{\Sigma}_x\mathbf{H}^T\mathbf{R}^{-1} \tag{12.28}$$

with $\boldsymbol{\Sigma}_x$, the covariance matrix of the estimation error $\Delta\mathbf{x} = \mathbf{x} - \hat{\mathbf{x}}$:

$$\boldsymbol{\Sigma}_x = \mathbb{E}\left[\,\Delta\mathbf{x}\;(\Delta\mathbf{x})^T\,\right] = \left(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\right)^{-1} \tag{12.29}$$

Note the following important properties of matrix $\mathbf{Q}$:

$$\text{rank}(\mathbf{Q}) = m - n = k, \quad \mathbf{I} > \mathbf{Q} \geqslant 0, \quad \mathbf{Q}^2 = \mathbf{Q} \tag{12.30}$$

For two Hermitian matrices $\mathbf{A}$ and $\mathbf{B}$, we say $\mathbf{A} \geq \mathbf{B}$ if $\mathbf{A} - \mathbf{B}$ is a positive semi-definite matrix.

In the absence of bad data, the measurement residual vector is distributed:

$$\mathcal{N}\left(0, \mathbf{Q}\mathbf{R}\mathbf{Q}^T\right) = \mathcal{N}(0, \mathbf{Q}\mathbf{R}) \tag{12.31}$$

The detection criteria currently used in this section are:

$$\text{the weighted residual vector } \mathbf{r}_W = \sqrt{\mathbf{R}^{-1}}\mathbf{r} \tag{12.32}$$

$$\text{the normalized residual vector } \mathbf{r}_N = \sqrt{\mathbf{D}^{-1}}\mathbf{r} \text{ with } \mathbf{D} = \text{diag}(\mathbf{Q}\mathbf{R}) \tag{12.33}$$

$$\text{the quadratic cost function } J(\hat{\mathbf{x}}) = \mathbf{r}^T\mathbf{R}^{-1}\mathbf{r} = \mathbf{r}_W^T\mathbf{r}_W \tag{12.34}$$

The detection of bad data is based on a hypothesis testing with the two hypotheses $\mathcal{H}_0$ and the alternative $\mathcal{H}_1$ where:

$$\begin{aligned} \mathcal{H}_0 &: \text{ no bad data are present} \\ \mathcal{H}_1 &: \mathcal{H}_0 \text{ is not true, i.e., there are bad data} \end{aligned} \tag{12.35}$$

Denoting by P_e the probability of rejecting $\mathcal{H}_0$ when $\mathcal{H}_0$ is actually true and P_d the probability of accepting $\mathcal{H}_1$ when $\mathcal{H}_1$ is true (probability of detection), the hypothesis testing consists in comparing $J(\hat{\mathbf{x}})$, $\left|r_{W_i}\right|$ or $\left|r_{N_i}\right|$ with a "detection threshold" γ, which depends on P_e. For example, considering the normalized residuals, one has:

- accepting $\mathcal{H}_0$ if $\left|r_{N_i}\right| < \gamma$, $i = 1, 2, \ldots, m$;
- rejecting $\mathcal{H}_0$ (and hence accept $\mathcal{H}_1$) otherwise.

12.12 Improved Bad Data Detection

Interesting links between outlier identification and ℓ_0-(pseudo)-norm minimization are presented in [592] and [593] under the Bayesian and the frequentist frameworks, respectively. Recently, ℓ_1-norm-based methods have been devised [592–594].

While the primary purpose of the installation of PMUs is not related to state estimation, widespread placement of PMUs presents an opportunity to improve state estimation. Capability of a state estimator to detect bad data is directly related to the measurement configuration. Bad data associated with critical measurements cannot be detected. Transforming critical measurements into redundant measurements requires adding extra measurements at strategic locations. See [595] for details. Besides, the authors in [596] used PMUs to transform the critical measurements into redundant measurements such that the bad measurements can be detected by the measurement residual testing.

12.13 Cyber-Attacks

As a complex cyber-physical system spanning a large geographical area, the power grid inevitably faces challenges in terms of cyber security. With more data acquisition and two-way communication required for the future grid, enhancing cyber security is of paramount importance.

12.14 Line Outage Detection

Although phasor measurement units (PMUs) have become increasingly widespread throughout power networks, the buses monitored by PMUs still constitute a very small percentage of the total number of system buses. Our problem is to derive useful information from PMU data in spite of this limited coverage. In particular, we can exploit known system topology information, together with PMU phasor angle measurements, to detect system line outages.

It is possible to use PMU data, even when coverage is extremely limited, to detect system events. The knowledge of topology changes outside of the local control area could be obtained by using data that is currently available on the North American power grid. The algorithm of [597] can provide a robust way to increase operator awareness of line statuses throughout an electric interconnection.

Bibliographical Remarks

We have drawn material freely from [145] in scattered places throughout the whole of this chapter.

We followed [569] for the modeling of the PMUs in Section 12.1.

In Section 12.6, we followed [598, 599] for the classical scheme for state estimation.

In Section 12.11, we followed [598,600] for the classical results on bad data detection.

Section 12.14 was taken from [597].

13

False Data Injection Attacks against State Estimation

This chapter gives an exhaustive treatment of false data injection attacks in the context of state estimation. It is well known that maintaining cyber security is the most important task facing engineers and researchers. We use false data injection to attack against state estimation.

Triggered by the seminar work of Liu and Ning and Reiter [601], a new line of research in power system security has focused on cyber intrusion related to intelligent electronic devices, such as remote terminal units, phasor measurement units and meters.

As an important module in the modern power control system, the state estimation program uses the measurements from intelligent electronic devices to estimate state variables like voltage angles and magnitudes at each bus in a power system. Statistical techniques successfully identify and remove obvious bad data from state estimation procedures. As state estimation cleans the data, this process also prevents the bad data from being stored in databases for future use.

13.1 State Estimation

Monitoring power flows and voltages in a power system is critical to system reliability. To ensure that a power system continues to operate even when some components fail, a number of meters are used to monitor system components and report readings to the control center, which estimates the state of power system variables according to these meter measurements. The state variables of interest include bus voltage angle and magnitudes.

The state estimation problem is to estimate power system state variables $\mathbf{x} = \left(x_1, x_2, \ldots, x_n\right)^T \in \mathbb{R}^{n\times 1}$ based on the meter measures $\mathbf{z} = \left(z_1, z_2, \ldots, z_m\right)^T \in \mathbb{R}^{m\times 1}$, where n and m are integers. The measurements errors (or uncertainties) are modeled as $\mathbf{e} = \left(e_1, e_2, \ldots, e_m\right)^T \in \mathbb{R}^{m\times 1}$. As a result, the state variables are related to the measurements through the following model

$$\mathbf{z} = \mathbf{h}\left(\mathbf{x}\right) + \mathbf{e} \tag{13.1}$$

where $\mathbf{h}\left(\mathbf{x}\right) = \left(h_1\left(x_1, x_2, \ldots, x_n\right), \ldots, h_m\left(x_1, x_2, \ldots, x_n\right)\right) \in \mathbb{R}^{m\times 1}$ and $h_i\left(x_1, x_2, \ldots, x_n\right)$ is a function of $x_1, x_2, \ldots, x_n$. The state estimation problem is to find an estimate $\hat{\mathbf{x}}$ of $\mathbf{x}$ that is the best fit of the measurement $\mathbf{z}$ according to (13.1).

Smart Grid using Big Data Analytics: A Random Matrix Theory Approach, First Edition.
Robert C. Qiu and Paul Antonik.

For state estimation using the DC power flow model, (13.1) can be represented by a **linear** regression model

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{e} \tag{13.2}$$

where $\mathbf{H} = \left(h_{i,j}\right) \in \mathbb{R}^{m \times b}$, a Jacobian matrix, while $\mathbf{Hx}$ is a vector of m linear functions linking measurements to states.

Of course, the linear model of (13.2) is much easier to handle than the nonlinear model of (13.1). Three basic statistical criteria are commonly used in state estimation [602]: the *maximum likelihood criterion*, the *weighted least-square criterion*, and the *minimum variance criterion*. When meter error is assumed to be normally distributed with zero mean, these three criteria lead to an identical estimator with the following matrix solution

$$\hat{\mathbf{x}} = \left(\mathbf{H}^T\mathbf{W}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{W}\mathbf{H} \tag{13.3}$$

where $\mathbf{W}$ is a diagonal matrix whose elements are reciprocals of the variances of meter errors. That is

$$\mathbf{W} = \begin{pmatrix} \sigma_1^{-2} & & 0 \\ & \ddots & \\ 0 & & \sigma_n^{-2} \end{pmatrix} \tag{13.4}$$

where σ_i^2 is the variance of the i-th meter ($1 \leq i \leq n$).

Bad measurement detection (also called bad data detection) may be introduced due to various reasons such as meter failures and malicious attack. Techniques for bad measurements detection have been developed to protect state estimation [602, 603].

The measurement residual $\mathbf{z} - \mathbf{H}\hat{\mathbf{x}}$ and its ℓ_2-norm $\|\mathbf{z} - \mathbf{H}\hat{\mathbf{x}}\|$ are used to detect the presence of bad measurements. Specifically, $\|\mathbf{z} - \mathbf{H}\hat{\mathbf{x}}\|$ is compared with a threshold τ, and the presence of bad measurements is assumed if

$$\|\mathbf{z} - \mathbf{H}\hat{\mathbf{x}}\| > \tau \tag{13.5}$$

The selection of τ is a key issue. Assume that all the state variables are mutually independent and the meter errors follow the normal distribution. It is known that $\|\mathbf{z} - \mathbf{H}\hat{\mathbf{x}}\|^2$ follows a $\chi^2(m-n)$-distribution with the degree of freedom $m-n$. According to [602], τ can be determined through a hypothesis test with a significance level α. Thus, $\|\mathbf{z} - \mathbf{H}\hat{\mathbf{x}}\|^2 \geqslant \alpha$ indicates the presence of bad measurements, with the probability of a false alarm being α.

Recently, the focus in bad measurement processing is on the improvement of the robustness using phasor measurement units (PMUs) [591, 595, 596, 604]. See Section 12.11 for the background on bad data detection.

It seems that these approaches targeting arbitrary, interacting bad measurements (e.g. [598, 600, 605, 606]) can also defeat the malicious ones injected by attackers, because such malicious measurements are indeed arbitrary, interacting bad measurements. A fundamental flaw of these approaches is that all of them use the same method, i.e., $\|\mathbf{z} - \mathbf{H}\hat{\mathbf{x}}\|^2 \geqslant \tau$, to detect the presence of bad measurements. In the next section, following [601], we will show that an attacker can systematically bypass this detection method, and thus all existing approaches.

13.2 False Data Injection Attacks

We assume that there are m meters that provide m measurements $z_1, \ldots, z_m$ and there are n state variables $x_1, \ldots, x_n$. The relationship between these m meter measurements and n state variables can be characterized by a *linear* $m \times n$ matrix $\mathbf{H}$, as given by (13.2). Of course the nonlinear relation between $\mathbf{z}$ and $\mathbf{x}$ can be addressed later. In general, the matrix $\mathbf{X}$ of a power system is a *constant* matrix determined by the topology and line impedances of the system. In [603], how the control center constructs $\mathbf{H}$ is illustrated.

Key assumptions for the model are:

- The meter measurements $\mathbf{z}$ and the state variables $\mathbf{x}$ are related by the linear matrix $\mathbf{H}$, i.e., given by (13.2).
- The matrix $\mathbf{H}$ is constant.
- The attacker can have access to the matrix $\mathbf{H}$ of the target power system.
- The attacker can inject malicious measurements into compromised meters to undermine the state estimation process.

13.2.1 Basic Principle

Let $\mathbf{z} + \mathbf{a}$ represent the vector of observed measurements that may contain malicious data, where $\mathbf{z} = \left(z_1, \ldots, z_m\right)^T$ is the vector of original measurements and $\mathbf{a} = \left(a_1, \ldots, a_m\right)^T$ is the malicious data added to the original measurements. We refer to $\mathbf{a}$ as an attack vector. Let $\hat{\mathbf{x}}_{\text{bad}}$ and $\hat{\mathbf{x}}$ denote the estimates of $\mathbf{x}$ using the malicious measurements $\mathbf{z} + \mathbf{a}$ and the original measurements $\mathbf{z}$, respectively. $\hat{\mathbf{x}}_{\text{bad}}$ can be represented as $\hat{\mathbf{x}} + \mathbf{b}$, where $\mathbf{b}$ is a nonzero vector of length n. $\mathbf{b}$ reflects the estimation error injected by the attacker.

Theorem 13.2.1 Suppose the original measurements $\mathbf{z}$ can pass the bad measurement detection defined by (13.5). Then, the malicious measurements $\mathbf{z} + \mathbf{a}$ can also pass the bad measurement detection (13.5), if $\mathbf{a}$ is a linear combination of the column vectors of $\mathbf{H}$, i.e., $\mathbf{a} = \mathbf{Hb}$.

Proof. As $\mathbf{z}$ can pass the detection (13.5), we have

$$\|\mathbf{z} - \mathbf{H}\hat{\mathbf{x}}\| \leqslant \tau, \tag{13.6}$$

where τ is the detection threshold. Recall that $\hat{\mathbf{x}}_{\text{bad}}$ can be represented as $\hat{\mathbf{x}} + \mathbf{b}$, where $\mathbf{b}$ is a nonzero vector of length n. Considering the condition that $\mathbf{a} = \mathbf{Hb}$, so $\mathbf{a}$ is a linear combination of the column vectors $\mathbf{h}_1, \ldots, \mathbf{h}_n$ of $\mathbf{H}$, then the resulting ℓ_2-norm of the measurement residual satisfies

$$\begin{aligned} &\|(\mathbf{z} + \mathbf{a}) - \mathbf{H}(\hat{\mathbf{x}} + \mathbf{b})\| \\ &= \|\mathbf{z} - \mathbf{H}\hat{\mathbf{x}} + (\mathbf{a} - \mathbf{Hb})\| \\ &= \|\mathbf{z} - \mathbf{H}\hat{\mathbf{x}}\| \leqslant \tau \end{aligned} \tag{13.7}$$

Thus, the ℓ_2-norm of the measurement residual of $\mathbf{z} + \mathbf{a}$ is less than the threshold τ. This means $\mathbf{z} + \mathbf{a}$ can also pass the bad measurement detection. □

Observing the derivation in (13.7), it is sufficient to consider the relaxed requirement with high probability for $\varepsilon > 0$

$$\begin{aligned} \|(\mathbf{z}+\mathbf{a}) - \mathbf{H}(\mathbf{H}\hat{\mathbf{x}}+\mathbf{b})\| &= \|\mathbf{z} - \mathbf{H}\hat{\mathbf{x}} + (\mathbf{a}-\mathbf{H}\mathbf{b})\| \\ &= \|\mathbf{z} - \mathbf{H}\hat{\mathbf{x}}\| + \|\mathbf{a}-\mathbf{H}\mathbf{b}\| \\ &\leqslant \tau + \varepsilon\tau = (1+\varepsilon)\tau \end{aligned} \tag{13.8}$$

provided that with high probability

$$\|\mathbf{a}-\mathbf{H}\mathbf{b}\| \leqslant \varepsilon\tau \tag{13.9}$$

Liu, Ning and Reiter observe in [601] that if there exists a nonzero k-sparse $\mathbf{a}$ for which $\mathbf{a} = \mathbf{H}\mathbf{b}$ for some $\mathbf{b}$, then

$$\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{a} + \mathbf{e} = \mathbf{H}(\mathbf{x}+\mathbf{b}) + \mathbf{e}$$

Thus as a deterministic quantity, $\mathbf{x}$ is observationally equivalent to $\mathbf{x}+\mathbf{b}$. No detector can distinguish $\mathbf{x}$ from $\mathbf{x}+\mathbf{b}$, so we will call an attack vector $\mathbf{a}$ *unobservable* if it has the form $\mathbf{a} = \mathbf{H}\mathbf{b}$.

It is unlikely that random bad data $\mathbf{a}$ will satisfy $\mathbf{a} = \mathbf{H}\mathbf{b}$. But an adversary can synthesize its attack vector to satisfy the unobserved condition.

13.3 MMSE State Estimation and Generalized Likelihood Ratio Test

A power system is composed of a collection of buses, transmission lines, and power flow meters. We adopt a graph-theoretical model for such a system. The power system is modeled as an undirected graph $(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of buses, and $\mathcal{E}$ is the set of transmission lines. Each line connect two meters, so each element $e \in \mathcal{E}$ is an unordered pair of buses in $\mathcal{V}$.

The control center receives measurements from various meters deployed throughout the system, from which it performs state estimation. The goal of state estimation is to recover the full state of the system: the voltage and phase of every bus in the network. Meters come in two types: transmission line-flow meters, which measure the power flow through a single transmission line, and bus injection meters, which measure the total outgoing flow on all transmission lines connected to a single bus. So each meter is associated with either a bus in $\mathcal{V}$ or a line in $\mathcal{E}$. We allow for the possibility of multiple meters on the same bus or line.

The $\mathbf{H}$ matrix in (13.12) arises from the graph theoretical model as follows. For each transmission line $(i,j) \in \mathcal{E}$, the DC power flow through i to j is $B_{ij}(x_i - x_j)$, where B_{ij} is the susceptance of the transmission line (i,j). We may also write this power flow as $\mathbf{h}_{ij}\mathbf{x}$, where

$$\mathbf{h}_{ij} = \begin{bmatrix} 0\cdots 0 & \underbrace{B_{ij}}_{\text{i-th element}} & 0\cdots 0 & \underbrace{-B_{ij}}_{\text{j-th element}} & 0\cdots 0 \end{bmatrix} \tag{13.10}$$

If a meter measures the flow through the transmission line connecting buses i and j, the associated row of $\mathbf{H}$ is therefore obtained by $\mathbf{h}_{ij}$. A bus injection meter measures the

total power flow on all lines incident to a particular node. The row of $\mathbf{H}$ associated with a meter on bus i is therefore given by

$$\sum_{j:(i,j)\in\mathcal{E}} \mathbf{h}_{ij} \tag{13.11}$$

The graph-theoretical model for the power system gives the following DC power flow model, a linearized version of the AC flow model:

$$\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{a} + \mathbf{e} \tag{13.12}$$

where

$$\mathbf{e} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}_e\right), \quad \mathbf{a} \in \mathcal{A}_k = \left\{\mathbf{a} \in \mathbb{R}^m : \|\mathbf{a}\|_0 \leqslant k\right\}$$

Here, $\mathbf{z} \in \mathbb{R}^m$ is the vector power flow measurements, $\mathbf{x} \in \mathbb{R}^n$ the system state, $\mathbf{e} \in \mathbb{R}^m$ the Gaussian measurement noise with zero mean and covariance matrix $\mathbf{\Sigma}_e$, and vector $\mathbf{a}$ is malicious data injected by an adversary. Below we assume that the adversary can at most control k meters: $\mathbf{a}$ is a vector with at most k nonzero entries ($\|\mathbf{a}\|_0 \leqslant k$). A vector $\mathbf{a}$ is said to have sparsity k if $\|\mathbf{a}\|_0 \leqslant k$ where $\|\mathbf{a}\|_0$ represents the ℓ_0-norm, equal to the number of the nonzero entries of $\mathbf{a}$.

We assume that the adversary has access to network parameters $\mathbf{H}$ and is able to coordinate attacks from different meters. These assumptions, and the assumption that the adversary may choose any set of k meters it likes, give the adversary more power than perhaps possible in practice, which is a well adopted practice when analyzing security.

13.3.1 A Bayesian Framework and MMSE Estimation

In a Bayesian framework, the state variables are random vectors with Gaussian distribution $\mathcal{N}\left(\boldsymbol{\mu}_x, \mathbf{\Sigma}_x\right)$. We assume that, in practice, the mean vector $\boldsymbol{\mu}_x$ and covariance matrix $\mathbf{\Sigma}_x$ can be estimated from historical data. By subtracting the mean vector from the data, we can assume without loss of generality that $\boldsymbol{\mu}_x = \mathbf{0}$.

In the absence of an attack, where $\mathbf{a} = \mathbf{0}$ in (13.12), $(\mathbf{x}, \mathbf{z})$ are jointly Gaussian. The minimum mean square error (MMSE) estimator of the state vector is a linear vector given by

$$\hat{\mathbf{x}}(\mathbf{z}) = \arg\min_{\hat{\mathbf{x}}} \mathbb{E}\left(\|\mathbf{x} - \hat{\mathbf{x}}(\mathbf{z})\|^2\right) = \mathbf{K}\mathbf{z} \tag{13.13}$$

where

$$\mathbf{K} = \mathbf{\Sigma}_x \mathbf{H}^T \left(\mathbf{\Sigma}_x \mathbf{H}^T \mathbf{\Sigma}_x + \mathbf{\Sigma}_e\right)^{-1} \tag{13.14}$$

The minimum mean-square error, in the absence of an attack, is given by

$$\mathcal{E}_0 = \min_{\hat{\mathbf{x}}} \mathbb{E}\left(\|\mathbf{x} - \hat{\mathbf{x}}(\mathbf{z})\|^2\right) = \mathrm{Tr}\left(\mathbf{\Sigma}_x - \mathbf{K}\mathbf{H}\mathbf{\Sigma}_x\right)$$

If an adversary injects malicious data $\mathbf{a} \in \mathcal{A}_k$, and the control center is unaware of the attack, then the state estimator defined in (13.13) is no longer the true MMSE estimator (in the presence of attack); the estimator $\hat{\mathbf{x}}(\mathbf{z}) = \mathbf{K}\mathbf{z}$ ignores the possibility of attack and

it will cause a higher mean square error (MSE). In particular, it is easy to see that the MSE in the presence of $\mathbf{a}$ is given by

$$\mathcal{E}_0 + \|\mathbf{K}\mathbf{a}\|_2^2 \tag{13.15}$$

The second term in (13.15) represents the impact on the estimator from a particular attack vector $\mathbf{a}$. To increase the MSE at the state estimator, the adversary necessarily has to increase the "energy" of the attack, which increases the probability of being detected at the control center.

13.3.2 Statistical Model and Attack Hypotheses

Following [592], we present a formulation of the detection problem at the control center. We assume a Bayesian model where the state variables are random with a multivariate Gaussian distribution $\mathbf{x} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}_x\right)$. The detection model, on the other hand, is not Bayesian in the sense that we do not assume any prior probability of the attack, nor do we assume any statistical model for the attack vector $\mathbf{a}$.

Under the observation model (13.12), we consider the following composite binary hypothesis:

$$\mathcal{H}_0 \colon \mathbf{a} = \mathbf{0} \quad \text{versus} \quad \mathcal{H}_1 \colon \mathbf{a} \in \mathcal{A}_k \backslash \{\mathbf{0}\} \tag{13.16}$$

Given observation $\mathbf{z} \in \mathbb{R}^m$, we wish to design a detector $\Lambda \colon \mathbb{R}^m \to \{0, 1\}$ with $\Lambda(\mathbf{z}) = 1$ indicating a detection of attack $\mathcal{H}_1$ and $\Lambda(\mathbf{z}) = 0$ the null hypothesis $\mathcal{H}_0$.

An alternative formulation is based on the extra MSE $\|\mathbf{K}\mathbf{a}\|_2^2$ at the state estimator. See (13.15). In particular, we may want to distinguish, for $\|\mathbf{a}\|_0 \leqslant k$

$$\mathcal{H}_0 \colon \|\mathbf{K}\mathbf{a}\|_2^2 \leqslant \tau \quad \text{versus} \quad \mathcal{H}_1 \colon \|\mathbf{K}\mathbf{a}\|_2^2 > \tau \tag{13.17}$$

where τ is the detection threshold.

13.3.3 Generalized Likelihood Ratio Detector with ℓ_1-Norm Regularization

For the hypothesis test given by (13.16), the uniformly most powerful test does not exist. We propose a detector based on the generalized likelihood ratio test (GLRT). We note that if we have multiple measurements under the same $\mathbf{a}$, the GLRT is asymptotically optimal in the sense that it offers the fatest decay rate of missing detection probability [607].

The distribution of the measurement $\mathbf{z}$ under the two hypotheses differ only in their means:

$$\begin{aligned}
\mathcal{H}_0 &: \mathbf{z} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}_z\right) \\
\mathcal{H}_1 &: \mathbf{z} \sim \mathcal{N}\left(\mathbf{a}, \mathbf{\Sigma}_z\right), \quad \mathbf{a} \in \mathcal{A}_k \backslash \{\mathbf{0}\}
\end{aligned}$$

where $\mathbf{\Sigma}_z = \mathbf{H}\mathbf{\Sigma}_x\mathbf{H}^T + \mathbf{\Sigma}_e$. The GLRT is given by

$$L(\mathbf{z}) \triangleq \frac{\max\limits_{\mathbf{a}\in\mathcal{A}_k} f(\mathbf{z}|\mathbf{a})}{f(\mathbf{z}|\mathbf{a}=\mathbf{0})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \tau \tag{13.18}$$

where $f(\mathbf{z}|\mathbf{a})$ is the Gaussian density function with mean $\mathbf{a}$ and covariance matrix $\mathbf{\Sigma}_z$, and the threshold is chosen from considering the null hypothesis at a certain false alarm rate. This is equivalent to

$$\min_{\mathbf{a}\in\mathcal{A}_k} \mathbf{a}^T\mathbf{\Sigma}_z^{-1}\mathbf{a} - 2\mathbf{z}^T\mathbf{\Sigma}_z^{-1}\mathbf{a} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \tau \tag{13.19}$$

Then the GLRT reduces to solving

$$\begin{aligned} &\text{minimize} \quad \mathbf{a}^T\boldsymbol{\Sigma}_z^{-1}\mathbf{a} - 2\mathbf{z}^T\boldsymbol{\Sigma}_z^{-1}\mathbf{a} \\ &\text{subject to} \quad \|\mathbf{a}\|_0 \leqslant k \end{aligned} \tag{13.20}$$

which is a nonconvex optimization since the ℓ_0-norm is nonconvex. It is well known that (13.20) can be approximated by a convex optimization:

$$\begin{aligned} &\text{minimize} \quad \mathbf{a}^T\boldsymbol{\Sigma}_z^{-1}\mathbf{a} - 2\mathbf{z}^T\boldsymbol{\Sigma}_z^{-1}\mathbf{a} \\ &\text{subject to} \quad \|\mathbf{a}\|_1 \leqslant \nu \end{aligned} \tag{13.21}$$

where the ℓ_1-norm constraint is a heuristic for the sparsity of $\mathbf{a}$. The constant ν needs to be adjusted until the solution involves an $\mathbf{a}$ with sparsity k. This requires solving (13.21) several times. A similar approach is used in [608].

13.3.4 Classical Detectors with MMSE State Estimation

Two classical bad data detectors [574, 609] are based on the residual error $\mathbf{r} = \mathbf{z} - \mathbf{H}\hat{\mathbf{x}}$ resulting from the MMSE state estimator.

The first is the $J(\hat{\mathbf{x}})$ detector, defined as

$$\mathbf{r}^T\boldsymbol{\Sigma}_e^{-1}\mathbf{r} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \tau \tag{13.22}$$

The second is the largest normalized residue (LNR) test given by

$$\max_i \frac{|r_i|}{\sigma_{r_i}} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \tau \tag{13.23}$$

where σ_{r_i} is the standard deviation of the i-th residual error r_i. This test can be viewed as the test on the ℓ_∞-norm of the measurement residual, which is normalized so that each element has unit variance.

The asymptotic optimality of the GLRT detector implies a better performance of GLRT over the above two detectors when the sample size is large.

13.3.5 Optimal Attacks for the MMSE and the GLRT Detector

We assume that the attacker has the prior knowledge that the MMSE and the GLRT detector are used by the control center. We also assume that the attacker can choose k meters arbitrarily in which for the attacker to inject malicious data.

The attacker has two conflicting objectives: maximizing the MSE by choosing the best data injection $\mathbf{a}$ versus avoiding being detected by the control center. Using (13.23), we can formulate the problem as

$$\begin{aligned} &\underset{\mathbf{a}\in\mathcal{A}_k}{\text{maximize}} \quad \|\mathbf{K}\mathbf{a}\|_2^2 \\ &\text{subject to} \quad \Pr(\Lambda(\mathbf{z}) = 1 \,|\mathbf{a}) \leqslant \beta \end{aligned} \tag{13.24}$$

or equivalently

$$\begin{aligned} &\text{minimize} \quad \Pr(\Lambda(\mathbf{z}) = 1 \,|\mathbf{a}) \\ &\text{subject to} \quad \|\mathbf{K}\mathbf{a}\|_2^2 \geqslant C \\ &\qquad\qquad \|\mathbf{a}\|_0 = k \end{aligned} \tag{13.25}$$

Due to the lack of analytical expressions for the detection error probability $\Pr(\Lambda(\mathbf{z}) = 1 \,|\mathbf{a})$, the solution of (13.24) and (13.25) is very difficult. We present a heuristic for $\Pr(\Lambda(\mathbf{z}) = 1 \,|\mathbf{a})$, which allows us to get the approximation solution.

Given the naive MMSE state estimator $\hat{\mathbf{x}} = \mathbf{Kz}$ (13.13) and (13.14), the estimation residual error is given by

$$\mathbf{r} = \mathbf{Gz}, \quad \mathbf{G} = \mathbf{I} - \mathbf{HK} \tag{13.26}$$

Inserting the measurement model, we obtain

$$\mathbf{r} = \mathbf{GHx} + \mathbf{Ga} + \mathbf{Ge}$$

where $\mathbf{Ga}$ is the only term from the attack. From (13.15), the damage in MSE done by injecting $\mathbf{a}$ is $\|\mathbf{Ka}\|_2^2$. So we can consider the equivalent problems:

$$\begin{aligned} \text{maximize} \quad & \|\mathbf{Ka}\|_2^2 \\ \text{subject to} \quad & \|\mathbf{Ga}\|_2^2 \leqslant \eta \\ & \|\mathbf{a}\|_0 = k \end{aligned} \tag{13.27}$$

or equivalently,

$$\begin{aligned} \text{minimize} \quad & \|\mathbf{Ga}\|_2^2 \\ \text{subject to} \quad & \|\mathbf{Ka}\|_2^2 \geqslant C \\ & \|\mathbf{a}\|_0 = k \end{aligned} \tag{13.28}$$

After some procedures, solving the optimal attack vector $\mathbf{a}$ for the above two formulations amounts to a standard generalized eigenvalue problem. See [592] for details.

The state estimation is used to set prices and calculating payment. As malicious attacks can change the state estimation significantly, it is natural to consider the impact of an attack on the electricity market [592].

13.4 Sparse Recovery from Nonlinear Measurements

State estimation for nonlinear electrical power networks is considered for bad data detection. The problem is formulated in terms of sparse recovery from nonlinear measurements. In the presence of bad data vector $\mathbf{v}$, the nonlinear model (13.1) is rewritten as

$$\mathbf{z} = \mathbf{h}(\mathbf{x}) + \mathbf{e} + \mathbf{a} \tag{13.29}$$

where $\mathbf{h}(\mathbf{x})$ is a set of n general functions, which may be linear or nonlinear, and $\mathbf{e}$ is the vector of the additive measurement noise. Here we assume that $\mathbf{e}$ is an n-dimensional vector with i.i.d. zero mean Gaussian elements of variance σ^2. We also assume that $\mathbf{a}$ is a vector with at most k nonzero entries, and the nonzero entries can take arbitrary real-number values. The sparsity k of gross errors reflects the nature of bad data because generally only a few faulty sensing results are present or an adversary party may control only a few malicious meters.

In the absence of bad data, it is well known that the standard least square (LS) method can be used to suppress the effect of observation noise on state estimations. Here, we

consider the nonlinear LS method, where we try to find a vector $\mathbf{x}$ that minimizes the least-square error:

$$\text{minimize} \quad \|\mathbf{y} - \mathbf{h}(\mathbf{x})\|_2 \tag{13.30}$$

However, the LS method generally only works well in the absence of bad data $\mathbf{a}$. If the magnitudes of bad data are large, the estimation result can be very far from the true state.

Bad data detection in power grids can be viewed as a sparse error detection problem, which shares mathematical structures similar to sparse recovery problems in compressed sensing. Since $\mathbf{h}(\mathbf{x})$ is a nonlinear mapping instead of a linear mapping in the compressed sensing, the problem here is different from that of the conventional compressed sensing.

13.4.1 Bad Data Detection for Linear Systems

For a special case of $\mathbf{h}(\mathbf{x}) = \mathbf{H}\mathbf{x}$, (13.29) becomes

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{a} + \mathbf{e} \tag{13.31}$$

where $\mathbf{x}$ is an $m \times 1$ signal vector ($m < n$), $\mathbf{H}$ is an $n \times m$ matrix, $\mathbf{a}$ is a sparse error vector with at most k nonzero elements, and $\mathbf{e}$ is a noise vector with $\|\mathbf{e}\|_2 \leqslant \varepsilon$.

We solve the following optimization problem involving optimization variables $\mathbf{x}$ and $\mathbf{z}$. The state estimation $\hat{\mathbf{x}}$ is the optimizer value for $\mathbf{x}$

$$\begin{array}{ll} \underset{\mathbf{x},\mathbf{z}}{\text{minimize}} & \|\mathbf{y} - \mathbf{H}\mathbf{x} - \mathbf{z}\|_1 \\ \text{subject to} & \|\mathbf{z}\|_2 \leqslant \varepsilon \end{array} \tag{13.32}$$

A subspace in $\mathbb{R}^n$ satisfies the *almost Euclidean* property [610, 611] for a constant $\alpha \leq 1$, if

$$\alpha\sqrt{n}\|\mathbf{x}\|_2 \leqslant \|\mathbf{x}\|_1$$

holds true for every $\mathbf{x}$ in the subspace.

We denote the part of any vector $\mathbf{w}$ over any index set K as $\mathbf{w}_K$.

Theorem 13.4.1 ([594]) Suppose that the minimum nonzero singular value of $\mathbf{H}$ is $\sigma_{\min}$. Let $C > 1$ be a real number, and suppose that every vector $\mathbf{w}$ in range of the matrix $\mathbf{H}$ satisfies $C\|\mathbf{w}_K\|_1 \leqslant \|\mathbf{w}_{\overline{K}}\|_1$ for any subset $K \subseteq \{1, 2, \ldots, n\}$ with cardinality $|K| \leqslant k$, where k is an integer, and $\overline{K} = \{1, 2, \ldots, n\} \backslash K$. We also assume the subspace generated by $\mathbf{H}$ satisfies the almost Euclidean property for a constant $\alpha \leq 1$. Then the solution $\hat{\mathbf{x}}$ to (13.32) satisfies

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \leqslant \frac{2(C+1)}{\sigma_{\min}\alpha(C-1)}\varepsilon \tag{13.33}$$

Proof. The proof is taken from [594]. Suppose that one optimal solution pair to (13.32) is $(\hat{\mathbf{x}}, \hat{\mathbf{z}})$. Since $\|\hat{\mathbf{z}}\|_2 \leqslant \varepsilon$, we obtain $\|\hat{\mathbf{z}}\|_1 \leqslant \sqrt{n}\|\hat{\mathbf{z}}\|_2 \leqslant \sqrt{n}\varepsilon$.

Since $\mathbf{x}$ and $\mathbf{z} = \mathbf{e}$ are feasible for (13.32) and $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{a} + \mathbf{e}$, then

$$\begin{aligned} \|\mathbf{y} - \mathbf{H}\hat{\mathbf{x}} - \hat{\mathbf{z}}\|_1 &= \|\mathbf{H}(\mathbf{x} - \hat{\mathbf{x}}) + \mathbf{a} + \mathbf{e} - \hat{\mathbf{z}}\|_1 \\ &\leqslant \|\mathbf{H}(\mathbf{x} - \mathbf{x}) + \mathbf{a} + \mathbf{e} - \mathbf{e}\|_1 = \|\mathbf{a}\|_1 \end{aligned}$$

Applying the triangle inequality to $\|\mathbf{H}(\mathbf{x}-\hat{\mathbf{x}})+\mathbf{a}+\mathbf{e}-\hat{\mathbf{z}}\|_1$, we have

$$\|\mathbf{H}(\mathbf{x}-\hat{\mathbf{x}})+\mathbf{a}\|_1-\|\mathbf{e}\|_1-\|\hat{\mathbf{z}}\|_1\leqslant\|\mathbf{a}\|_1$$

Denoting $\mathbf{H}(\mathbf{x}-\hat{\mathbf{x}})$ by $\mathbf{w}$, since $\mathbf{a}$ is supported on a set K with cardinality $|K|\leqslant k$, by the triangle inequality for ℓ_1-norm,

$$\|\mathbf{a}\|_1-\left\|\mathbf{w}_K\right\|_1+\left\|\mathbf{w}_{\overline{K}}\right\|_1-\|\mathbf{e}\|_1-\|\hat{\mathbf{z}}\|_1\leqslant\|\mathbf{a}\|_1$$

Thus we have

$$-\left\|\mathbf{w}_K\right\|_1+\left\|\mathbf{w}_{\overline{K}}\right\|_1\leqslant\|\hat{\mathbf{z}}\|_1+\|\mathbf{e}\|_1\leqslant 2\sqrt{n}\varepsilon \tag{13.34}$$

With $C\left\|\mathbf{w}_K\right\|_1\leqslant\left\|\mathbf{w}_{\overline{K}}\right\|_1$ in the assumption, we know

$$\frac{C+1}{C-1}\|\mathbf{w}\|_1\leqslant-\left\|\mathbf{w}_K\right\|_1+\left\|\mathbf{w}_{\overline{K}}\right\|_1$$

Combining this with (13.34), we have

$$\frac{C+1}{C-1}\|\mathbf{w}\|_1\leqslant 2\sqrt{n}\varepsilon$$

By the almost Euclidean property $\alpha\sqrt{n}\|\mathbf{x}\|_2\leqslant\|\mathbf{x}\|_1$, we have

$$\|\mathbf{w}\|_2\leqslant\frac{2(C+1)}{\alpha(C-1)}\varepsilon \tag{13.35}$$

By the definition of singular values

$$\sigma_{\min}\|\mathbf{x}-\hat{\mathbf{x}}\|_2\leqslant\|\mathbf{H}(\mathbf{x}-\hat{\mathbf{x}})\|_2=\|\mathbf{w}\|_2 \tag{13.36}$$

so combing this with (13.35), we obtain the desired result

$$\|\mathbf{x}-\hat{\mathbf{x}}\|_2\leqslant\frac{2(C+1)}{\sigma_{\min}\alpha(C-1)}\varepsilon$$

□

In the absence of sparse errors, the decoding error bound using the standard LS method satisfies [612] $\|\mathbf{x}-\hat{\mathbf{x}}\|_2\leqslant\frac{1}{\sigma_{\min}}\varepsilon$.

13.4.2 Bad Data Detection for Nonlinear Systems

Theorem 13.4.2 ([594]) Let $\mathbf{y}=\mathbf{h}(\mathbf{x})+\mathbf{a}$. A state $\mathbf{x}$ can be recovered correctly from any error with $\|\mathbf{a}\|_0\leqslant k$ from solving the optimization

$$\underset{\mathbf{x}}{\text{minimize}}\quad\|\mathbf{y}-\mathbf{h}(\mathbf{x})\|_0 \tag{13.37}$$

if and only if for any $\mathbf{x}^\star\neq\mathbf{x}$, $\left\|\mathbf{h}(\mathbf{x})-\mathbf{h}\left(\mathbf{x}^\star\right)\right\|_0\geqslant 2k+1$.

Theorem 13.4.3 ([594]) Let $\mathbf{y}=\mathbf{h}(\mathbf{x})+\mathbf{a}$. A state $\mathbf{x}$ can be recovered correctly from any error with $\|\mathbf{a}\|_0\leqslant k$ from solving the optimization

$$\underset{\mathbf{x}}{\text{minimize}}\quad\|\mathbf{y}-\mathbf{h}(\mathbf{x})\|_1 \tag{13.38}$$

if and only if for any $\mathbf{x}^\star \neq \mathbf{x}$, $\left\| \left(\mathbf{h}\left(\mathbf{x}\right) - \mathbf{h}\left(\mathbf{x}^\star\right) \right)_K \right\|_1 < \left\| \left(\mathbf{h}\left(\mathbf{x}\right) - \mathbf{h}\left(\mathbf{x}^\star\right) \right)_{\overline{K}} \right\|_1$, where K is the support of the error vector $\mathbf{a}$.

Direct ℓ_0 and ℓ_1 minimization may be computationally costly because ℓ_0 norm and nonlinear $\mathbf{h}(\cdot)$ may lead to nonconvex optimization problems. Reference [594] proposes a computationally efficient iterative sparse recovery algorithm for the general setting of the additive noise $\mathbf{e}$.

13.5 Real-Time Intrusion Detection

Recent research in power-system security has focused entirely on cyber intrusion related to intelligent electronic devices (IEDs) like remote terminal units (RTUs), phasor measurement units (PMUs), and meters. These attacks are referred to as malicious data injection attacks. The research in [613] defines a new class of cyber attacks to power systems—malicious modification of network data stored in an accessible database—which is different from the research on malicious data injection attacks.

Network data stored in databases is also vulnerable to cyber attack. These cyber attacks are different from previously researched data integrity attacks in the sense that these physical transmission line data do not depend on the measurements from IEDs.

Bibliographical Remarks

The material in Section 13.1 is taken from [601].

The material in Section 13.2 is taken from [601].

In Section 13.3, we take material from [592, 614]. Relevant work is [601] and [608]. Bad data detection is a classical problem that is part of the original formulation of state estimation [574]. The formulation in Section 13.4 is taken from [594].

See [615] for a survey. It is a key task in smart grid to send the readings of smart meters to an access point (AP) in a wireless manner. Compressed sensing can be used [616].

14

Demand Response

Smart grid is primarily envisioned as a quantum leap in harnessing communication and information technologies to enhance grid reliability and to enable integration of various smart grid resources such as renewable resources, demand response, electric storage, and electric transportation.

14.1 Why Engage Demand?

The concept of smart grid started with the notion of advanced metering infrastructure to improve demand-side management, energy efficiency, and a self-healing electrical grid to improve supply reliability and respond to natural disasters or malicious sabotage. One emerging paradigm shift is the increased and bidirectional interaction between wholesale markets/transmission operations and retail markets/distribution operations. The expected profusion of demand response, renewable resources, and distributed generation and storage at the distribution/retail level has direct implications on the operation of the transmission system and the wholesale energy markets. Enabling technologies, such as enhancements in the communication and information technologies, make it possible to turn these new resources into useful controllable products for wholesale market and transmission system operators.

For smart grid, the efforts are categorized into the following trends: (i) reliability; (ii) renewable resources; (iii) demand response; (iv) electric storage; (v) electric transportation.

System reliability has always been a major focus area for the design and operation of modern grids. Demand response and electric storage resources are necessary to address the economics of the grid and are perceived to support grid reliability by mitigating peak demand and load variability. Electric transportation resources are deemed helpful for meeting environmental targets and can be used to mitigate load variability. Balancing the diversity of the characteristics of these resource types presents challenges in maintaining grid reliability.

Meeting these reliability challenges while effectively integrating the above resources requires a quantum leap in harnessing communication and information technologies. Wide-area monitoring and control is important. This involves gathering data from and controlling a large region of the grid through the use of time-synchronized phasor measurement units (PMUs). Big data analytics are promising in this direction.

Smart Grid using Big Data Analytics: A Random Matrix Theory Approach, First Edition.
Robert C. Qiu and Paul Antonik.

Demand and resource forecasting is usually done at a macroscopic level, such as control area and load zone. As the need for more discrete and intelligent local control increases, better forecast at the local level will be required for demand and distributed resources. One approach is to use forecasting agents throughout the grid to communicate and access required data and information to produce more accurate load and generation models throughout the system. We need failproof geographically and temporally coordinated hierarchical monitoring and control actions over timescales ranging from milliseconds to operational planning horizon.

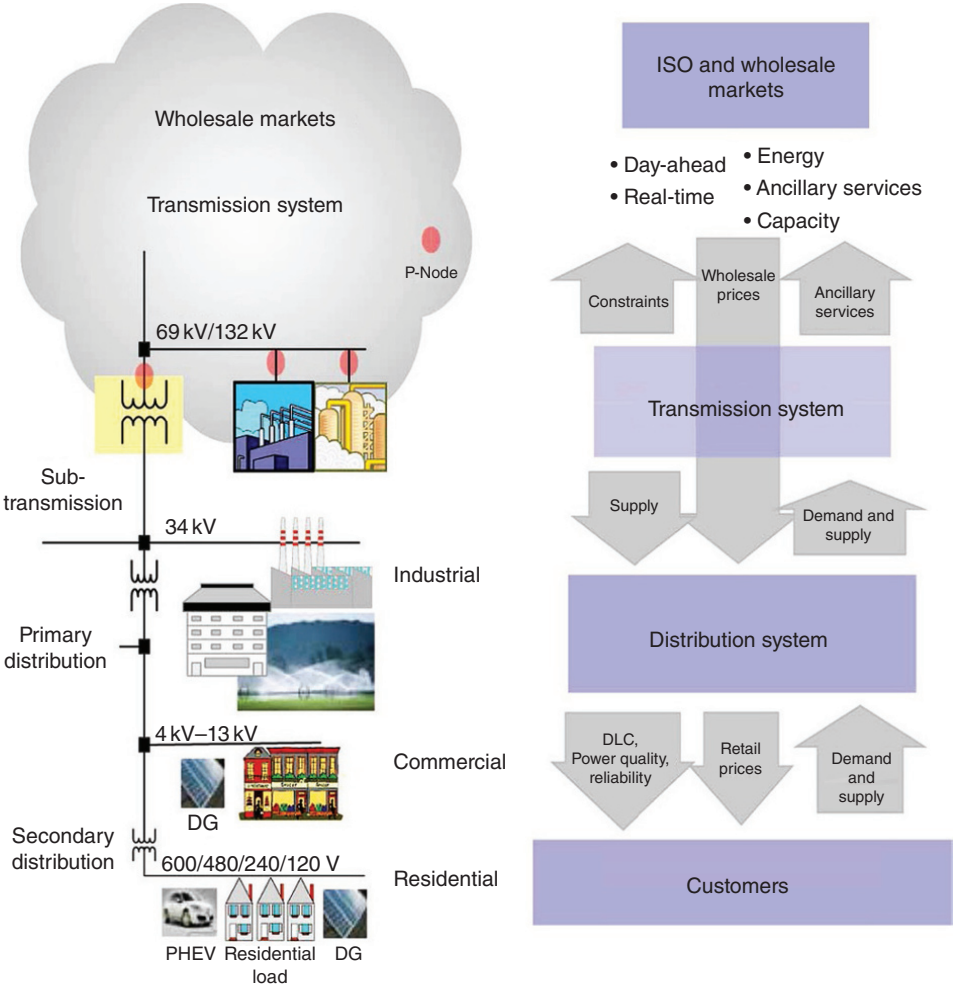


Figure 14.1 Demand response connectivity and information flow. Source: Reproduced from [617] with permission.

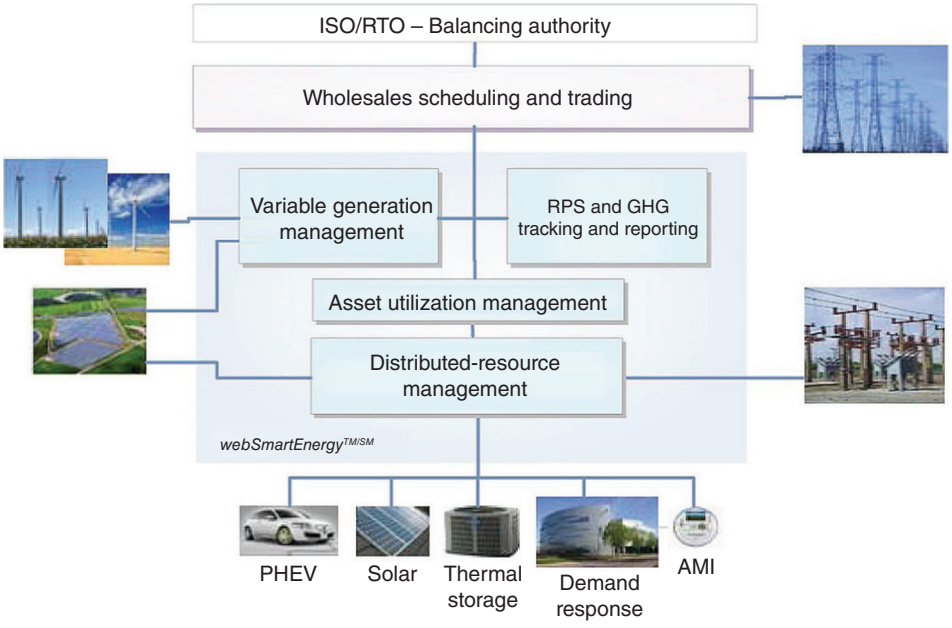


Figure 14.2 Interaction of demand response, variable generation, and storage. Source: Reproduced from [617] with permission.

In Figure 14.1, demand response connectivity and information flow are illustrated. In Figure 14.2, the interaction of demand response, variable generation, and storage is illustrated. See Section 9.4 for more illustrations.

We illustrate the general wholesale electricity market scenario shown in Figure 14.3, where each retailer/utility serves a number of end users. The real-time pricing information, reflecting the wholesale prices, is informed by the retailer to the users over a digital communication infrastructure, for example, a local area network (LAN).

This is the right moment to make links with the rest of the book on big data analytics. In Figure 14.1 and Figure 14.2, it is vividly shown that the grid is distributed. Information flow is behind the control of this distributed grid. Massive amounts of data are generated in this system for big physical data. Mathematically, these data are modeled as a matrix valued time series $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$, where $\mathbf{X}_i$, $i = 1, \ldots, n$ are large random matrices of $N \times T$. For N sensors, we can zoom in the time period (in milliseconds) of $t = 1, \ldots, T$ for one basic building block. We deal with a total of time window (in milliseconds) $t = T$, $2T, \ldots, nT$. We are especally interested in the asymptotic regime when both N and T go large with the fixed concentration $c = N/T$. For example, the values of N and T are in the order of $100 - 10000$. Examples of sensors include: (i) PMUs; (ii) smart meters; (iii) antennas. The examples are almost endless.

In Figure 14.3, we can apply the large random matrices similarly to Figure 14.1 and Figure 14.2.

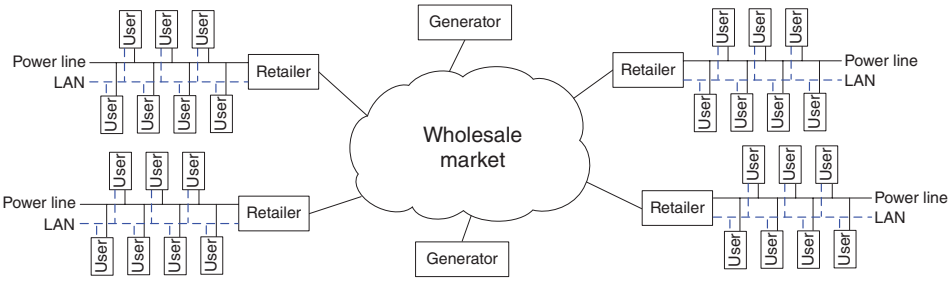


Figure 14.3 A simplified illustration of the wholesale electricity market formed by multiple generators and several regional retail companies. Each retailer provides electricity for a number of users. Retailers are connected to the users via local area networks which are used to announce real-time prices to the users. Source: Reproduced from [548] with permission of the IEEE.

14.2 Optimal Real-time Pricing Algorithms

Electricity is currently provided through an infrastructure consisting of utility companies, power plants, and transmission lines, which serve millions of customers. The dependency of almost all parts of industry and different aspects of our life on electrical energy makes this massive infrastructure a strategic entity.

Currently, electricity consumption is not efficient in most buildings (e.g., due to poor thermal isolation). This results in the waste of a large amount of natural resources, since most of the electricity consumption occurs in buildings. Besides, new types of demand such as plug-in hybrids will potentially double the average household load. For the above reasons, there is a need to develop new methods for demand-side management (DSM).

There is a wide range of DSM techniques such as voluntary load management programs and direct load control. Smart pricing is known as one of the most common tools that can encourage users to consume wisely and more efficiently. Users are often willing to improve the insulation conditions of their buildings or try to shift the energy consumption schedule of their high-load household appliances to off-peak hours. It is important to understand the real-time interactions among subscribers and the energy provider and real-time pricing algorithms for the future smart grid.

The problem of demand-side management is essentially a problem of economy—understanding the scarcity of resources and finding ways to allocate them. In economy, optimization plays a central role. By analogy, in the context of demand side management, optimization plays a fundamental role. The traditional work horse for optimization is linear programing (LP). In recent years, convex optimization such as semidefinite programing (SDP) has been the standard tool for researchers and engineers who work in the field of communications and signal processing. The formulated convex problems can be solved using convex programming techniques such as the interior point method [377]. Standard software packages such as CVX (written in MATLAB) [378] and CVXOPT (written in Python) [379] can be used. As we know, once the problem is formulated in terms of convex optimization, the rest is essentially the technical details. Linear programing is a special case of convex optimization. We often take advantage of the special structure of LP, rather than treating it as a convex

optimization problem. The main motivation of the work horse like SDP is to extend the LP.

We now give several examples to illustrate how the problems are formulated in terms of the above framework of convex optimization.

Example 14.2.1 (linear programing [618]) A simple LP algorithm to be integrated into the energy management system of a household or a small business. Via bidirectional communication with the electricity supplier, such an algorithm allows maximizing the consumer utility or minimizing its energy cost. The interaction takes place on a hourly basis using a rolling window algorithm to consider the energy consumption throughout the day. □

Example 14.2.2 (distributed concave optimization [619]) The formulated problem is a concave maximization problem and can be solved using convex programming techniques such as the interior point method in a central fashion. As this central formulation depends on the exact utility function of users that may not known, they formulate the problem in a distributed manner. The proposed algorithm is based on utility maximization. It can be implemented in a distributed manner to maximize the aggregate utility of all users and minimize the cost to the energy provider while keeping the total power consumption below the generating capacity. □

Example 14.2.3 (energy consumption scheduling [620]) The problem is formulated in terms of convex optimization. An optimal, autonomous, and incentive-based energy consumption scheduling algorithm can be used to balance the load among residential subscribers who share a common energy source. This kind of algorithm is designed to be implemented in energy consumption scheduling (ECS) devices inside smart meters in a smart-grid infrastructure. A simple pricing and billing model provides the incentives for the subscribers encouraging them to actually use the ECS devices and run the proposed distributed algorithm in order to be charged less. □

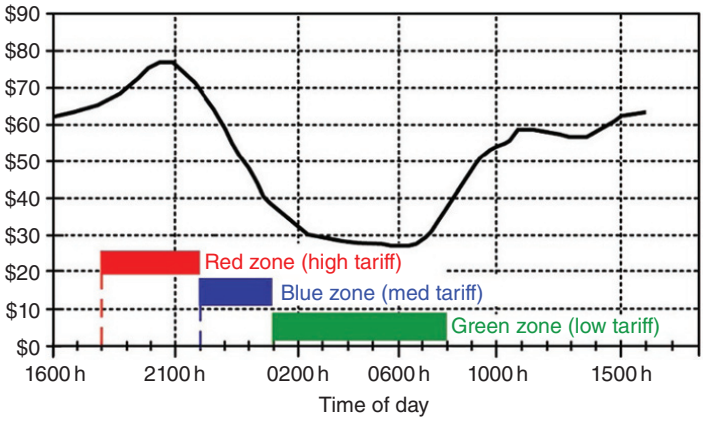


Figure 14.4 Daily residential load curve. Source: Reproduced from [621] with permission.

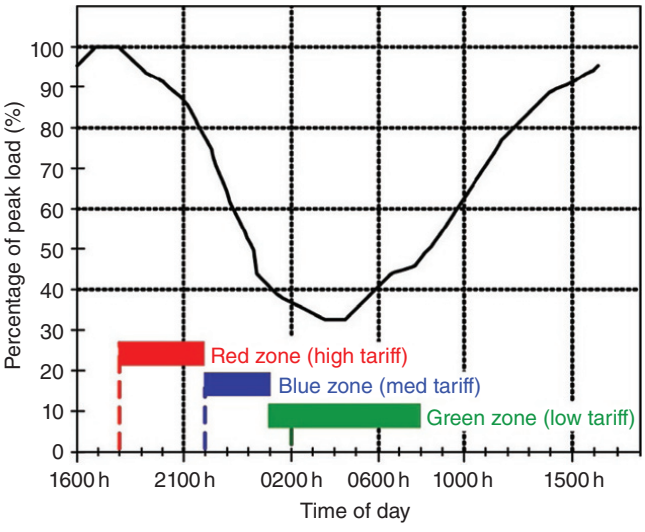


Figure 14.5 Subscription options of charging time zones for PEV owners and variable short-term market energy pricing. Source: Reproduced from [621] with permission.

14.3 Transportation Electrification and Vehicle-to-Grid Applications

Plug-in electric vehicles (PEVs) are growing in popularity as more efficient low-emission alternatives to the conventional fuel-based automobiles.

A new real-time smart load management (RT-SLM) approach for the coordination of PEV charging is used to improve the security and reliability of smart grids by minimizing voltage deviations, overloads, and power losses that would otherwise be impaired by random uncoordinated PEV charging. The random and unpredictable nature of PEV activity in a domestic household situation calls for a fast and adaptable real-time coordination strategy.

A real-time smart load management (RT-SLM) control strategy appropriately considers random plug-ins of PEVs and utilizes a maximum sensitivities selection (MSS) optimization approach to minimize system losses.

In Figure 14.4, the daily residential load curve is shown. In Figure 14.5, subscription options of charging time zones for PEV owners and variable short-term market energy pricing are shown.

14.4 Grid Storage

Large-scale energy storage system is an important part of the smart grid. It is the sixth part of the electric system besides generation, transmission, substation, distribution, and users.

Bibliographical Remarks

In Chapter 14, we draw material from [589, 617–621, 623, 624].
The grid storage is addressed in [625].

Part III

Communications and Sensing

15

Big Data for Communications

In [39], we treat a communication system as a big data system and model the massive amount of data with the aid of large random matrices. This book follows the same viewpoint—see Figure 1.6 for illustration.

15.1 5G and Big Data

We argue that the fifth-generation (5G) wireless communication network should be aware of big data [39, 40]. From the viewpoint of data processing, it is natural to model the massive amount of data using (large) random matrices. The theme of the 5G network (five disruptive technologies) belongs in a unified framework of big data.

15.2 5G Wireless Communication Networks

The 5G network is expected to be standardized around 2020. Compared with the 4G network, the 5G network should achieve 1000 times the system capacity, 10 times the spectral efficiency, energy efficiency and data rate (i.e., peak data rate of 10 Gb/s for low mobility and peak data rate of 1 Gb/s for high mobility), and 25 times the average cell throughput.

A proposed architecture for 5G wireless communication networks is shown in [141]. There are five disruptive technologies.

- **Device-centric architectures**. The base-station-centric architecture of cellular systems may change in 5G. We present device-centric architectures.
- **Millimeter wave (mmWave)**. While spectrum has become scarce at microwave frequencies, it is plentiful in the mmWave realm. There is an mmWave "gold rush." Although far from being fully understood, mmWave technologies have already been standardized for short-range services (IEEE 802.11ad) and deployed for niche applications such as small-cell backhaul.
- **Massive MIMO.** Massive multiple-input multiple-output (MIMO) proposes utilizing a very high number of antennas to multiplex messages for several devices on each time-frequency resource, focusing the radiated energy toward the intended directions while minimizing intracell and intercell interference. Massive MIMO may require major architectural changes, particularly in the design of macro base stations, and it may also lead to new types of deployments.

Smart Grid using Big Data Analytics: A Random Matrix Theory Approach, First Edition.
Robert C. Qiu and Paul Antonik.

- **Smarter devices**. 2G-3G-4G cellular networks were built under the design premise of having complete control at the infrastructure side. We argue that 5G systems should drop this design assumption and exploit intelligence at the device side within different layers of the protocol stack, for example by allowing device-to-device (D2D) connectivity or exploiting smart caching at the mobile side. While this design philosophy mainly requires a change at the node level (component change) it also has implications at the architectural level.
- **Native support for machine-to-machine (M2M) communication**. A native inclusion of M2M communication in 5G involves satisfying three fundamentally different requirements associated with different classes of low-data-rate services: support for a massive number of low-rate devices, sustaining a minimal data rate in virtually all circumstances, and very-low-latency data transfer. Addressing these requirements in 5G requires new methods and ideas at both the component and architectural levels, and such is the focus of a later section.

Wireless communication is becoming a commodity, just like electricity or water. A massive number of connected devices are supported by the 5G network. Whereas current systems typically operate with, at most, a few hundred devices per base station, some M2M services might require over 104 connected devices. Examples include metering, sensors, smart grid components, and other enablers of services targeting wide-area coverage.

15.3 Massive Multiple Input, Multiple Output

Massive MIMO is a promising technology for next-generation wireless systems. Its relevance to sm grid (which is the next-generation power grid) is clear. The central message of this section is to tie together the massive MIMO system and the large random matrices. At the heart of problems of massive MIMO lie the big data challenges.

15.3.1 Multiuser-MIMO System Model

Consider the uplink of a multiuser-MIMO system. The system has one base station equipped with an array of M antennas that receive data from K single-antenna users. The users transmit their data in the same time-frequency resource. The $M \times 1$ received vector at the base station is

$$\mathbf{y} = \sqrt{P_{\text{avg}}}\mathbf{G}\mathbf{x} + \mathbf{n} \tag{15.1}$$

where $\mathbf{G}$ represents the $M \times K$ channel matrix between the base station and the K users, i.e., $g_{mk} \triangleq [\mathbf{G}]_{mk}$ is the channel coefficient between the m-th antenna of the base station and the k-th user, $\sqrt{P_{\text{avg}}}\mathbf{x}$ is the vector of symbols simultaneously transmitted by the K users (the average transmitted power of each user is P_{avg}), and $\mathbf{n}$ is a vector of additive white, zero-mean Gaussian noise. We take the noise variance to be 1, to minimize notation, but without loss of generality. With this convention, P_{avg} has the interpretation of normalized "transmit" signal-to-noise ratio (SNR) and is therefore dimensionless. The model (15.1) also applies to wideband channels handled by OFDM over restricted intervals of frequency.

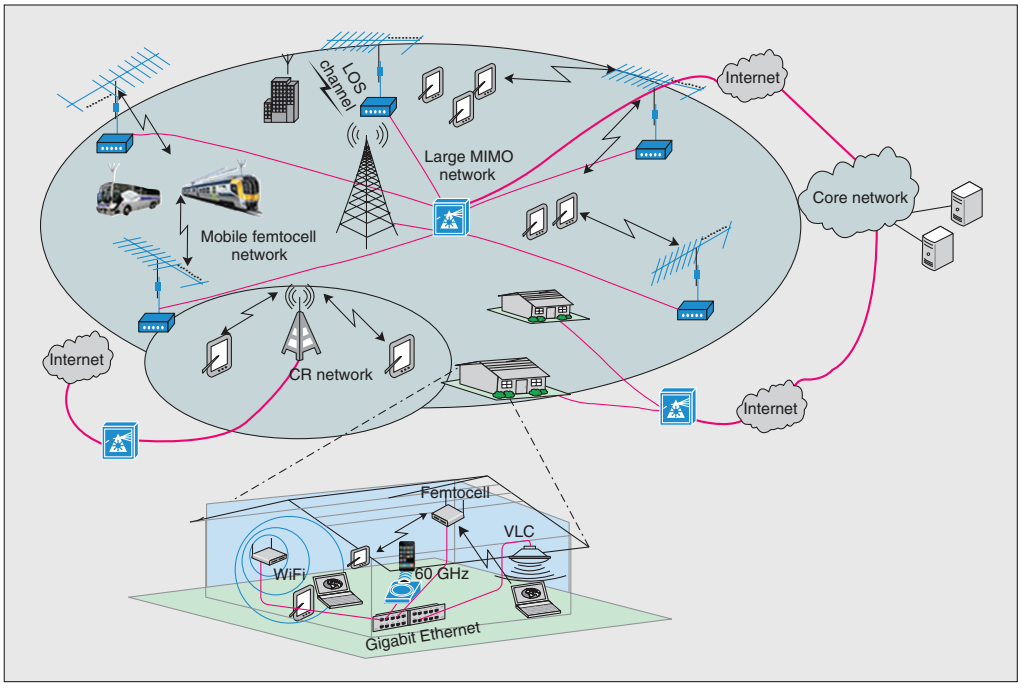


Figure 15.1 A proposed 5G heterogeneous wireless cellular architecture. Source: Reproduced with permission from [141].

The channel matrix $\mathbf{G}$ models independent fast fading, geometric attenuation, and log-normal shadow fading. The coefficient g_{mk} can be written as

$$g_{mk} = h_{mk}\sqrt{\beta_k}, \quad m - 1, 2, \ldots, M \tag{15.2}$$

where h_{mk} is the fast fading coefficient from the k-th user to the m-th antenna of the base station. $\sqrt{\beta_k}$ models the geometric attenuation and shadow fading, which is assumed to be independent over m and to be constant over many coherence time intervals and known a priori. Then, we have

$$\mathbf{G} = \mathbf{H}\mathbf{D}^{1/2} \tag{15.3}$$

where $\mathbf{H}$ is the $M \times K$ matrix of fast fading coefficients between the K users and the base station, i.e., $[\mathbf{H}]_{mk} = h_{mk}$, and $\mathbf{D}$ is a $K \times K$ diagonal matrix, where $[\mathbf{D}]_{kk} = \beta_k$.

15.3.2 Very Long Random Vectors

We review some results for random vectors from Cramer [626]. Let $\mathbf{x} \triangleq \left[X_1, \ldots, X_n\right]^T$ and $\mathbf{y} \triangleq \left[Y_1, \ldots, Y_n\right]^T$ be mutually independent $n \times 1$ random vectors whose elements are i.i.d. zero-mean random variables (RVs) with $\mathbb{E}\left|X_i\right|^2 = \sigma_x^2$, and $\mathbb{E}\left|Y_i\right|^2 = \sigma_y^2$, $i = 1, \ldots, n$. Then from the law of large numbers, we have

$$\frac{1}{n}\mathbf{x}^H\mathbf{x} \xrightarrow{a.s.} \sigma_x^2 \quad \text{and} \quad \frac{1}{n}\mathbf{x}^H\mathbf{y} \xrightarrow{a.s.} 0, \text{ as } n \to \infty \tag{15.4}$$

where $\xrightarrow{a.s.}$ denotes the almost sure convergence. Also, from the Lindeberg–Lévy central-limit theorem, we have

$$\frac{1}{\sqrt{n}}\mathbf{x}^H\mathbf{y} \xrightarrow{d} \mathcal{CN}\left(0, \sigma_x^2\sigma_y^2\right), \text{ as } n \to \infty \tag{15.5}$$

where $\xrightarrow{d}$ denotes convergence in distribution.

15.3.3 Favorable Propagation

Throughout the section, we assume that the fast fading coefficients, the elements of $\mathbf{H}$, are i.i.d. RVs with zero mean and unit variance. Then conditions (15.4) and (15.5) are satisfied with $\mathbf{x}$ and $\mathbf{y}$ being any two distinct columns of $\mathbf{G}$. In this case we obtain

$$\frac{1}{M}\mathbf{G}^H\mathbf{G} = \frac{1}{M}\mathbf{D}^{1/2}\mathbf{H}^H\mathbf{H}\mathbf{D}^{1/2} \approx \mathbf{D}, \quad M \gg K \tag{15.6}$$

and we say that we have *favorable propagation*. Obviously, if all fading coefficients are i.i.d. and zero mean, we have favorable propagation.

Remarkably, the total throughput (e.g., the achievable sum rate) of reverse link MU-MIMO is given by [627]

$$R_{\text{sum}} = \log_2 \det\left(\mathbf{I}_K + \frac{P_{\text{avg}}}{M}\mathbf{G}\mathbf{G}^H\right) \tag{15.7}$$

where $\mathbf{G}$ is a random matrix. This form of log-determinant of random matrices has been treated elsewhere in this book. See Section 15.3.6 and Section 15.4.1 for the connections.

To understand why favorable propagation is desirable, consider an $M \times K$ uplink (multiple-access) MIMO channel $\mathbf{H}$, where $M \geq K$, neglecting for now path loss and shadowing factors in $\mathbf{D}$. This channel can offer a sum-rate of

$$R_{\text{sum}} = \sum_{k=1}^{K} \log_2 \left(1 + P_{\text{avg}} \lambda_k^2\right) \tag{15.8}$$

where P_{avg} is the average power spent per terminal and $\left\{\lambda_k\right\}_{k=1}^{K}$ are the singular values of $\mathbf{H}$. If the channel matrix is normalized such that $\left|H_{ij}\right| \sim 1$ (where $\sim$ means equality of the order of magnitude), then $\sum_{k=1}^{K} \lambda_k^2 = \|\mathbf{H}\|_F^2 \approx MK$, $\|\cdot\|_F$ represents the Frobenius norm. Under this constraint the sum rate R_{sum} is bounded as

$$\log_2 \left(1 + MKP_{\text{avg}}\right) \leqslant R_{\text{sum}} \leqslant K \log_2 \left(1 + MP_{\text{avg}}\right) \tag{15.9}$$

The lower bound (left inequality) is satisfied with equality if $\lambda_1^2 = MK$ and $\lambda_2^2 = \cdots = \lambda_k^2 = 0$ and corresponds to a rank-one (line-of-sight) channel. The upper bound (right inequality) is achieved if $\lambda_1^2 = \cdots = \lambda_k^2 = M$. This occurs if the columns of $\mathbf{H}$ are mutually orthogonal and have the same norm, which is the case when we have favorable propagation.

Under this assumption of favorable propagation defined by (15.6), the base station could process its received signal by a matched-filter (MF)

$$\begin{aligned} \mathbf{G}^H \mathbf{y} &= \sqrt{P_{\text{avg}}} \mathbf{G}^H \mathbf{G} \mathbf{x} + \mathbf{G}^H \mathbf{n} \\ &\approx M \sqrt{P_{\text{avg}}} \mathbf{D} \mathbf{x} + \mathbf{G}^H \mathbf{n}, \quad \text{for } M >> K \end{aligned} \tag{15.10}$$

where we have used $\mathbf{G}^H \mathbf{G} \approx M\mathbf{D}$, which follows from (15.6).

Remarks and Links with Large Random Matrices

The bounds in (15.9) are too loose. Starting with (15.7) or (15.8), we can derive much tighter bounds. The key is the observation that $\mathbf{G}$ is a *large-dimensional random matrix*, and thus $\frac{1}{M}\mathbf{G}\mathbf{G}^H$ is the sample covariance matrix. The matrix type $\frac{1}{M}\mathbf{G}^H\mathbf{G} = \frac{1}{M}\mathbf{D}^{1/2}\mathbf{H}^H\mathbf{H}\mathbf{D}^{1/2}$ has been widely investigated in the random matrix literature. For example, see Section 15.3.6 and Section 15.4.1.

Roughly speaking, to obtain (15.6), the assumption of $M \gg K$ (this condition is required by the classical law of large numbers) is too strong, and can be greatly relaxed in the modern random matrix theory. This direction of research may be pursued.

The sample covariance matrix $\frac{1}{M}\mathbf{G}\mathbf{G}^H$ is of paramount significance to high-dimensional statistics. The sample covariance matrix $\frac{1}{M}\mathbf{G}\mathbf{G}^H$ also plays a central role here in obtaining both the bounds of the sum rate (15.9) and the precoding techniques (15.10) or (15.13).

Our departure point for statistical analysis usually starts with the sample covariance matrix $\frac{1}{n}\mathbf{Z}\mathbf{Z}^H$. Here $\mathbf{Z}$ is a complex $p \times n$ random matrix, and p and n go to infinity simultaneously: $p \to \infty$, $n \to \infty$, but their ratio is concentrated around c, $p/n \to c \in (0, \infty)$.

Let us first assume that the entries of $\mathbf{Z}$ are i.i.d with variance 1. Results on the global behavior of the eigenvalues of $\frac{1}{n}\mathbf{Z}\mathbf{Z}^H$ mostly concern the spectral distribution, that is

the $\frac{1}{p}\sum_{i=1}^{p}\delta_{\lambda_i}$, where δ denotes the Dirac measure. The spectral distribution converges, $n \to \infty$, $p \to \infty$ with $p/n \to c \in (0, 1]$, to a *deterministic* measure with density function

$$\frac{1}{2\pi c}\sqrt{(a-x)(b-x)}\mathbb{I}_{(a,b)}(x), \quad a = \left(1+\sqrt{c}\right)^2, \; b = \left(1-\sqrt{c}\right)^2$$

where $\mathbb{I}(x)$ is the indicator function. This is the so-called Marchenko–Pastur law.

The remarkable observation is when the size of random matrices is *sufficiently* large, we are able to exploit the unique phenomenon: deterministic spectral distribution is reached. The statistical properties of the entries of the large random matrix are so general and flexible.

15.3.4 Precoding Techniques

Assume that the base station has perfect knowledge of $\mathbf{G}$. Let $\mathbf{A}$ be an $M \times K$ linear detector matrix. By using the linear detector, the received signal vector $\mathbf{r}$ is separated into streams by multiplying it with $\mathbf{A}^H$ as follows

$$\mathbf{r} = \mathbf{A}^H\mathbf{y} \tag{15.11}$$

We have three conventional linear detectors: maximum ratio combining (MRC), zero forcing (ZF) and minimum mean-square error (MMSE)

$$\mathbf{A} = \begin{cases} \mathbf{G} & \text{for MRC} \\ \mathbf{G}\left(\mathbf{G}^H\mathbf{G}\right)^{-1} & \text{for ZF} \\ \mathbf{G}\left(\mathbf{G}^H\mathbf{G} + \frac{1}{P_{\text{avg}}}\mathbf{I}_K\right)^{-1} & \text{for MMSE} \end{cases} \tag{15.12}$$

From (15.1) and (15.11), the received vector after using the linear detector is given by

$$\mathbf{r} = \sqrt{P_{\text{avg}}}\mathbf{A}^H\mathbf{G}\mathbf{x} + \mathbf{A}^H\mathbf{n} \tag{15.13}$$

Let r_k and x_k be the k-th elements of the $K \times 1$ vectors $\mathbf{r}$ and $\mathbf{x}$, respectively. Then

$$r_k = \sqrt{P_{\text{avg}}}\mathbf{a}_k^H\mathbf{g}_k x_k + \sqrt{P_{\text{avg}}}\sum_{i=1, i\neq k}^{K}\mathbf{a}_i^H\mathbf{g}_i x_i + \mathbf{a}_i^H\mathbf{n} \tag{15.14}$$

where $\mathbf{a}_k$ and $\mathbf{g}_k$ are the k-th columns of the matrices $\mathbf{A}$ and $\mathbf{G}$, respectively. For a fixed-channel realization $\mathbf{G}$, the noise-plus-interference term is a random variable with zero mean and variance $P_{\text{avg}}\sum_{i=1, i\neq k}^{K}\left|\mathbf{a}_k^H\mathbf{g}_i\right|^2 + \|\mathbf{a}_k\|$, where $||\cdot||$ represents the Euclidean norm. By modeling this term as additive Gaussian noise independent of x_k, we can obtain a lower bound on the achievable rate. Assuming further that the channel is ergodic so that each codeword spans over a large (infinite) number of realizations of the fast-fading factor of $\mathbf{G}$, the ergodic achievable uplink rate of the k-th user is

$$R_{\text{sum}} = \mathbb{E}\log_2\det(1 + \text{SNR}) \tag{15.15}$$

where

$$\mathrm{SNR} = \frac{P_{\mathrm{avg}} \left| \mathbf{a}_k^H \mathbf{g}_k \right|^2}{P_{\mathrm{avg}} \sum_{i=1, i \neq k}^{K} \left| \mathbf{a}_k^H \mathbf{g}_i \right|^2 + \left\| \mathbf{a}_k \right\|} \tag{15.16}$$

When M grows large, $M \to \infty$, it follows from (15.6) that $\frac{1}{M}\mathbf{G}^H\mathbf{G} \to \mathbf{D}$, and hence the ZF and MMSE filters tend to that of the MRC. Thus, by using the law of large numbers, we can arrive at the same result for the ZF and MMSE receivers.

15.3.5 Downlink System Model

For each use of the channel the base station transmits an $M \times 1$ vector, $\mathbf{x}$ through its M antennas, and the K terminals collectively receive a $K \times 1$ vector, $\mathbf{y}$,

$$\mathbf{y} = \sqrt{\rho}\mathbf{G}^T\mathbf{x} + \mathbf{n} \tag{15.17}$$

where $\mathbf{n}$ is the $K \times 1$ vector of receiver noise whose components are independent and distributed as $CN(0, 1)$. The quantity ρ is proportional to the ratio of power to noise variance. The total transmit power is independent of the number of antennas,

$$\mathbb{E}\left(\|\mathbf{x}\|^2 \right) = 1 \tag{15.18}$$

where $||\cdot||$ represents the Euclidean norm of a vector.

The known capacity result for this channel, see, for example [628] and [629], assumes that the terminals as well as the base station know the channel $\mathbf{G}$. Let $\mathbf{\Gamma}$ [628, 629] be a diagonal matrix whose diagonal elements constitute a $K \times 1$ vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)^T$. To obtain the sum capacity requires performing a constrained optimization:

$$\begin{aligned} R_{\mathrm{sum}} &= \max_{\gamma_k} \log_2 \det\left(\mathbf{I}_M + \rho \mathbf{G}\mathbf{\Gamma}\mathbf{G}^H \right) \\ \text{subject to} \quad & \sum_{k=1}^{K} \gamma_k = 1, \quad \gamma_k \geqslant 0, \quad \forall k \end{aligned} \tag{15.19}$$

Under favorable propagation conditions (15.6) and a large excess of antennas, the sum capacity has a simple asymptotic form, By using the fundamental matrix identity (or *Sylvester determinant theorem*) that for all $\mathbf{A} \in \mathbb{C}^{p \times q}, \mathbf{B} \in \mathbb{C}^{q \times p}$

$$\det\left(\mathbf{I}_p + \mathbf{A}\mathbf{B} \right) = \det\left(\mathbf{I}_q + \mathbf{B}\mathbf{A} \right) \tag{15.20}$$

we have

$$\begin{aligned} R_{\mathrm{sum}} &= \max_{\gamma_k} \log \det_2 \left(\mathbf{I}_K + \rho \mathbf{\Gamma}^{1/2} \mathbf{G}\mathbf{G}^H \mathbf{\Gamma}^{1/2} \right) \\ &\approx \max_{\gamma_k} \log \det_2 \left(\mathbf{I}_K + M\rho\mathbf{\Gamma}\mathbf{D} \right) \\ &= \max_{\gamma_k} \sum_{k=1}^{K} \log_2 \left(1 + M\rho\gamma_k\beta_k \right) \end{aligned} \tag{15.21}$$

The significance of (15.20) is emphasized here in the context of random matrix theory [67, 345]. According to [67, p. 252], Percy Deift has half-jokingly termed this "the most important identity in mathematics." This formula is particular useful when computing

determinants of large matrices (or infinite-dimensional operators), as one can often use it to transform such determinants into much smaller determinants. In particular, the asymptotic behavior of $p \times p$ determinants as $p \to \infty$ can be converted via this formula into determinants of a fixed size (independent of p), which is often a more favorable situation to analyze.

We can use Roy's largest root test (Section 8.10) for the detection in (15.13) and (15.17).

15.3.6 Random Matrix Theory

So-called favorable propagation, or (15.6), plays a central role in the above asymptotic system analysis.

We have two asymptotic regimes: (i) K is fixed and $M \to \infty$, or $M \gg K$; (ii) $K \to \infty$, $M \to \infty$, but the ratio K/M tends to a fixed ratio $K/M \to c \in (0, \infty)$. Case (i) has been assumed in (15.6).

Now we consider case (ii), which belongs to the territory of random matrix theory [39]. The details can be found in the previous chapters. Our treatment of this new paradigm for massive MIMO is beyond the scope of this book and will be reported elsewhere.

- In Example 4.3.6, the MMSE receiver is treated using large random matrices.
- See also Section 7.8.1 for massive MIMO systems.
- In Example 3.6.3, the mutual information expression is valid for the massive MIMO analysis.
- The log-determinant of random matrices in 8.6 can be used for massive MIMO.

To some degree, the massive MIMO system may be viewed as a big data system in the sense of Table 3.1 in Section 3.2.

The goal of this example is to illustrate how the massive MIMO can be viewed to mimic the large CDMA systems.

Example 15.3.1 (multiuser CDMA systems) Consider a symbol synchronous direct sequence code division multiple access (DS-CDMA) system with K users. The discrete-time model for the received signal y in a symbol interval is

$$\mathbf{y} = \sum_{k=1}^{K} x_k \mathbf{s}_k + \mathbf{w} \tag{15.22}$$

where the x_k is the symbol transmitted by user k, $\mathbf{s}_k \in \mathbb{R}^N$ is the signature sequence of user k and $\mathbf{w}_k \in \mathbb{R}^N$ is the noise vector with mean zero and covariance matrix $\sigma^2\mathbf{I}$. We also assume that the symbol vector $\mathbf{x} = (x_1, \ldots, x_K)^T$ has a covariance matrix $\mathbf{P}$ where $\mathbf{P} = \text{diag}\left(P_1, \ldots, P_K\right)$ with P_k being the received power of user k, that is, $\mathbb{E}x_k^2 = P_k$ and that the symbol vector is uncorrelated with the noise. Putting $\mathbf{S} = \left(\mathbf{s}_1, \ldots, \mathbf{s}_K\right) \in \mathbb{R}^{N \times K}$, we rewrite (15.22) as

$$\mathbf{y} = \mathbf{S}\mathbf{x} + \mathbf{w} \tag{15.23}$$

Recall (15.1) that

$$\mathbf{y} = \sqrt{P_{\text{avg}}}\mathbf{G}\mathbf{x} + \mathbf{w} \tag{15.24}$$

where we replace $\mathbf{n}$ with $\mathbf{w}$ to use the same notation here. Using $\mathbf{G}$ as the precoder, we have

$$\mathbf{r} = \frac{1}{M}\mathbf{G}^H\mathbf{y} = \frac{1}{M}\mathbf{G}^H\mathbf{G}\mathbf{x} + \frac{1}{M}\mathbf{G}^H\mathbf{w} \tag{15.25}$$

or

$$\mathbf{r} = \mathbf{T}\mathbf{x} + \mathbf{w}' \tag{15.26}$$

where $\mathbf{T} = \frac{1}{M}\mathbf{G}^H\mathbf{G}$ and the filtered noise $\mathbf{w}' = \frac{1}{M}\mathbf{G}^H\mathbf{w}$ is also Gaussian. Clearly, the matrices $\mathbf{S}$ and $\mathbf{T}$ play similar roles. We can design the system to mimic the CDMA system, using this analogy.

The engineering goal is to demodulate the transmitted $x + k$ for each user. Assume that the receiver has already acquired the knowledge of the signature sequences. For user k, the linear minimum mean-square error (LMMSE) receiver generates an output in a form $\mathbf{a}_k^T\mathbf{y}$ where $\mathbf{a}_k$ is chosen to minimize the mean-squared error

$$\mathbb{E}\left|x_k - \mathbf{a}_k^T\mathbf{y}\right|^2 \tag{15.27}$$

The relevant performance measure is the signal-to-interference ratio (SIR) of the estimate, which is defined by

$$\beta_k = P_k\mathbf{s}_k^T\left(\mathbf{S}_k\mathbf{P}_k\mathbf{S}_k^T + \sigma^2\mathbf{I}\right)^{-1}\mathbf{s}_k, \quad k = 1, \ldots, K \tag{15.28}$$

where $\mathbf{S}_k$ and $\mathbf{P}_k$ are obtained from $\mathbf{S}$ and $\mathbf{P}$ by deleting the k-th column, respectively.

If signature sequences are modeled as being random, one may further proceed with the analysis using random matrix theory when the number of users K and the processing gain N approach infinity, that is, suppose

$$\mathbf{s}_k = \frac{1}{\sqrt{N}}\left(v_{1k}, \ldots, v_{Nk}\right)^T$$

for $k = 1, \ldots, K$, where $\left\{v_{ik}, i, k = 1, \ldots\right\}$ are independent and identically distributed (i.i.d.) random variables. ∎

Example 15.3.2 (capacity of the MIMO communication channel) Example 3.6.3, addresses massive MIMO capacity. We highlight some points. Denoting the number of transmitting antennas by M and the number of receiving antennas by N, the channel model is

$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{n} \tag{15.29}$$

where $\mathbf{s} \in \mathbb{C}^M$ is the transmitted vector, $\mathbf{y} \in \mathbb{C}$ is the received vector, $\mathbf{H} \in \mathbb{C}^{N\times M}$ is a complex matrix and $\mathbf{n} \in \mathbb{C}^N$ is the zero mean complex Gaussian vector with independent, equal variance entries.

Let $\mathbf{H}$ be an $n \times n$ Gaussian random matrix with complex, independent, and identically distributed entries of zero mean and unit variance. Given an $n \times n$ positive definite matrix $\mathbf{A}$, and a continuous function $f : \mathbb{R}^+ \to \mathbb{R}$ such that $\int_0^\infty e^{-\alpha t}|f(t)|^2 dt < \infty$ for every $\alpha > 0$, Tucci and Vega [54] find a new formula for the expectation

$$\mathbb{E}\left[\mathrm{Tr}\left(f\left(\mathbf{H}\mathbf{A}\mathbf{H}^H\right)\right)\right]$$

Taking $f(x) = \log(1 + x)$ gives another formula for the capacity of the MIMO communication channel, and taking $f(x) = (1 + x)^{-1}$ gives the MMSE achieved by a linear receiver. In this example, the matrix size n is arbitrary for $n \geq 2$. The 2×2 case is explicitly given in Example 3.6.3. ∎

Example 15.3.3 (distributed MIMO) Section 6.14 addresses Euclidean random matrices. This model may be applied to massive MIMO, where each antenna is viewed as a scattering center located at a random position $\mathbf{r}_i$, $i = 1, \ldots, N$. We are interested in the collective radiation from the region V containing N random located antennas. This is interesting when N is large, say $N = 10^4$. This is analogous with collective spontaneous emission in dense atomic systems consisting of N atoms [320]. One extension of the work in this section is to consider the impact of multipath on the eigenvalue distributions, because only the free-space Green's function is considered for the path with the line of sight (LOS) between the transmitter and the receiver.

For an arbitrary V, we have

$$\mathbf{A} = \mathbf{H}\mathbf{T}\mathbf{H}^H \tag{15.30}$$

The advantage of this representation lies in the separation of two different sources of complexity: the matrix $\mathbf{H}$ is random, but independent of the function f, whereas the matrix $\mathbf{T}$ depends on f but is not random. Often $\mathbf{T}$ is a Hermitian positive definite matrix.

Furthermore, if we assume that $\mathbb{E}H_{ij} = 0$, we readily find that H_{ij} are identically distributed random variables with zero mean and variance equal to $1/N$.

We will assume, in addition, that H_{ij} are independent Gaussian random variables. This assumption largely simplifies calculations but may limit applicability of our results at high densities of points ρ.

Now we can check that our model satisfies the conditions of Example 15.3.2, and thus use the results there to obtain the capacity. □

Example 15.3.4 (decentralized computing for eigenvalues) Large-data processing and analysis, often in real or near-real time, drives nearly every aspect of computing engineering. The ability to gather and analyze massive amounts of information will be a decisive factor for the fifth generation (5G) wireless communication system that must support a massive number of low-rate devices [140, 141]. A cognitive radio network (CRN) is also part of a proposed 5G heterogeneous wireless cellular architecture [140, 141]. To support this architecture, distributed computing is central to a large-scale cognitive radio network.

We require global solutions that optimize data collection, data modeling, and computing. The general solutions are not tractable analytically. We make two *fundamental assumptions*: (i) the massive data are modeled as large random matrices; (ii) algorithms only depend on eigenvalues of these random matrices.

We make assumption (ii) to simplify the algorithms required for distributed computing. Computing is critical to real-time applications such as detection and estimation. Often the eigenvalues of large random matrices need be computed in real time. As far as we are aware, all previous works (except a few papers) assume a centralized architecture, in which a fusion center gathers the signal samples received by all sensors, process the data, and forwards the decision back to all nodes. Such an architecture suffers from scalability issues and is not suitable for large-scale, multi-hop networks, where the fusion center may be many hops away from peripheral nodes.

For this reason, we seek a decentralized implementation of eigenvalue-based applications, so that computational effort is distributed across multiple sensor nodes communicating iteratively with neighbors (i.e., gossip algorithms) and the algorithm's statistics are computed locally by each node. We propose two solutions based on iterative eigenvalue algorithms—the power method and the Lanczos algorithm. In our large-scale network testbed at TTU, these algorithms can be implemented in a distributed fashion by applying distributed average consensus algorithms. The only relevant work is listed in [630–632]. □

15.4 Free Probability for the Capacity of the Massive MIMO Channel

See Section 5.8 for the basics of free probability. The deformed quarter-circle law has very important role in many practical fields. Consider the wireless MIMO system is defined as

$$\mathbf{y} = \sqrt{\gamma}\mathbf{H}\mathbf{x} + \mathbf{n}$$

The mutual information is derived as

$$\frac{1}{N} I(\gamma) = \int \log(1 + \gamma x) d\mathbb{P}_{\mathbf{H}\mathbf{H}^H}(x) \tag{15.31}$$

As an example in MIMO system, Rayleigh i.i.d. channel $\mathbf{H} \in \mathbb{C}^{N\times M}$ is a simple application of the deformed quarter circle law and the mutual information reads the following expression:

$$\begin{aligned}\lim_{N\to\infty} \frac{1}{N} \log\det\left(\mathbf{I} + \gamma\mathbf{H}\mathbf{H}^H\right) &= \lim_{N\to\infty} \frac{1}{N} \mathrm{Tr}\log\left(\mathbf{I} + \gamma\mathbf{H}\mathbf{H}^H\right) \\ &= \phi\left(\mathbf{I} + \frac{1}{\sigma^2}\mathbf{H}\mathbf{H}^H\right) \\ &= \int_{\left(1-\sqrt{c}\right)^2}^{\left(1+\sqrt{c}\right)^2} \log(1+\gamma x) \frac{\sqrt{4c - (x-1-c)^2}}{2\pi x} dx\end{aligned} \tag{15.32}$$

as $M, N \to \infty$ with the ratio $c = M/N$ fixed. It follows from (15.31) that

$$\frac{1}{N} I(\gamma) = \log\left(\frac{g(\gamma, c) - 2}{2\gamma g(\gamma, c)^c}\right) + \frac{g(\gamma, c)}{2\gamma} + c^2 \log c\gamma \tag{15.33}$$

where $g(x, y) \triangleq 1 + x - xy - \sqrt{(x+1)^2 + xy(xy + 2 - 2x)}$

15.4.1 Nonasymptotic Theory: Concentration Inequalities

The treatment in the literature belongs to the asymptotic regimes we mentioned above. There is the third regime: nonasymptotic theory using concentration of measure phenomenon. Qiu and Wicks [40] give an applied treatment of this topic. Due to space, we cannot go in depth here but we can highlight some points in the context of massive MIMO and big data.

- In Section 3.14, concentration of the spectral measure for large random matrices is studied in the form of $\mathrm{Tr}\, f(\mathbf{X})$ where $\mathbf{X}$ is a random matrix and f is a convex function.
- See Section 8.5 for some commonly used concentration inequalities.
- In Section 8.8.1, the eigenvalues apply to nonasymptotic, finite sample regimes. Combing this result with (15.8), we can obtain a new expression for the achievable sum rate.
- In Section 8.3, the eigenvalue bounds for expectation and variance are obtained. We can leverage these bounds by studying the functions of eigenvalues such as (15.8).

Let $\log^{\varepsilon}(x) = \log(\max(\varepsilon, x))$, and $\det^{\varepsilon}(\mathbf{X}) = \prod_i \max\left(\lambda_i(\mathbf{X}), \varepsilon\right)$ where $\mathbf{X}$ is a square Hermitian matrix.

Lemma 15.4.1 ([213]) Suppose that $n/p \in (0,1]$ is a fixed constant. Consider a real-valued random matrix $\mathbf{A} = \left[\xi_{ij}\right]_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p}$ where ξ_{ij} are jointly independent with zero mean and unit variance, and satisfy one of the following conditions:

- ξ_{ij} is almost surely bounded by a constant C.
- ξ_{ij} satisfies the logarithmic Sobolev inequality with uniformly bounded constant c_{LS}.

Then, for any $\varepsilon > 0$ and $t > \frac{4C\sqrt{\pi}}{\sqrt{n}(n+p)}$, there exists a constant $c > 0$ such that

$$\mathbb{P}\left(\left|\frac{1}{n}\log\det\left(\varepsilon + \frac{1}{n}\mathbf{A}\mathbf{A}^T\right) - \mathbb{E}\left(\frac{1}{n}\log\det\left(\varepsilon + \frac{1}{n}\mathbf{A}\mathbf{A}^T\right)\right)\right| > t\right) \leqslant 4\exp(-c\varepsilon^2 t^2 n^3) \tag{15.34}$$

and

$$\mathbb{P}\left(\left|\frac{1}{p}\log\det{}^{\varepsilon}\left(\frac{1}{n}\mathbf{A}\mathbf{A}^T\right) - \mathbb{E}\left(\frac{1}{p}\log\det{}^{\varepsilon}\left(\frac{1}{p}\mathbf{A}\mathbf{A}^T\right)\right)\right| > t\right) \leqslant 4\exp\left(-c\varepsilon^2 t^2 p^3\right) \tag{15.35}$$

Proof. Observe that the Lipschitz constant of $\log(\varepsilon + x)$ is upper bounded by $1/\varepsilon$ when $x \geq 0$. If ξ_{ij} is almost surely bounded by C and hence each entry of $\frac{1}{\sqrt{n}}\mathbf{A}$ is bounded by $\frac{1}{\sqrt{n}}C$, then applying Part (a) of Corollary 3.14.2 (which is [199, 1.8(a)]) leads to

$$\begin{aligned}&\mathbb{P}\left(\left|\frac{1}{n}\log\det\left(\varepsilon\mathbf{I} + \frac{1}{n}\mathbf{A}\mathbf{A}^T\right) - \mathbb{E}\left(\frac{1}{n}\log\det\left(\varepsilon\mathbf{I} + \frac{1}{n}\mathbf{A}\mathbf{A}^T\right)\right)\right| > \frac{n+p}{n}t\right)\\ &\quad\leqslant 4\exp\left(-\frac{\varepsilon^2}{4C^2}\left(t - \frac{2C\sqrt{\pi}}{\varepsilon\sqrt{n}(n+p)}\right)n(n+p)^2\right)\end{aligned}$$

Setting t to be a positive constant independent of n, we have for sufficiently large n that

$$\begin{aligned}&\mathbb{P}\left(\left|\frac{1}{n}\log\det\left(\varepsilon\mathbf{I} + \frac{1}{n}\mathbf{A}\mathbf{A}^T\right) - \mathbb{E}\left(\frac{1}{n}\log\det\left(\varepsilon\mathbf{I} + \frac{1}{n}\mathbf{A}\mathbf{A}^T\right)\right)\right| > \frac{n+p}{n}t\right)\\ &\quad\leqslant 4\exp\left(-\frac{\varepsilon^2 t^2}{4C^2}n^3\right)\end{aligned}$$

If ξ_{ij} satisfies the logarithmic Sobolev inequality with uniformly bounded constant c_{LS}, then the logarithmic Sobolev constant is bounded above by $\frac{1}{n}c_{LS}$, and hence Part (b) of Corollary 3.14.2 (which is [199, 1.8(b)]) leads to

$$\mathbb{P}\left(\left|\frac{1}{n}\log\det\left(\varepsilon\mathbf{I}+\frac{1}{n}\mathbf{A}\mathbf{A}^T\right)-\mathbb{E}\left(\frac{1}{n}\log\det\left(\varepsilon\mathbf{I}+\frac{1}{n}\mathbf{A}\mathbf{A}^T\right)\right)\right|>\frac{n+p}{n}t\right)$$
$$\leqslant 2\exp\left(-\frac{\varepsilon^2t^2(n+p)^2}{2c_{LS}}\right)$$

The proof is completed by observing that n/p is a given constant.

Given that the Lipschitz constant of the function $\log^{\epsilon}(x)$ is also $1/\varepsilon$, the concentration result for $\frac{1}{p}\log\det^{\epsilon}\left(\frac{1}{n}\mathbf{A}\mathbf{A}^T\right)$ follows with the same machinery. ∎

Now that we have established the concentration results for $\frac{1}{n}\log\det\left(\varepsilon\mathbf{I}+\frac{1}{n}\mathbf{A}\mathbf{A}^T\right)$, it remains to determine $\mathbb{E}\left(\frac{1}{n}\log\det\left(\varepsilon\mathbf{I}+\frac{1}{n}\mathbf{A}\mathbf{A}^T\right)\right)$. There are many ways. One is given above in Example 3.6.3. Here we show another approach.

Lemma 15.4.2 ([213]) Let $\mathbf{A}=\left[\xi_{ij}\right]_{1\leqslant i\leqslant n,1\leqslant j\leqslant p}$ be a real-valued random matrix such that ξ_{ij} are jointly independent with zero mean and unit variance. For any small constant $\varepsilon>0$, we have

$$\mathbb{E}\left(\frac{1}{n}\log\det\left(\varepsilon\mathbf{I}+\frac{1}{n}\mathbf{A}\mathbf{A}^T\right)\right)\leqslant\frac{1}{n}\log\mathbb{E}\left[\det\left(\varepsilon\mathbf{I}+\frac{1}{n}\mathbf{A}\mathbf{A}^T\right)\right]$$
$$\leqslant -1+O\left(\frac{\log n}{n}\right)+2\sqrt{\varepsilon} \tag{15.36}$$

Additionally, under Condition (a) or (b) of Lemma 15.4.1, $\mathbf{A}$ satisfies

$$\mathbb{E}\left(\frac{1}{n}\log\det\left(\varepsilon\mathbf{I}+\frac{1}{n}\mathbf{A}\mathbf{A}^T\right)\right)\geqslant -1-O\left(\frac{\log n}{n}\right) \tag{15.37}$$

15.5 Spectral Sensing for Cognitive Radio

Spectral sensing is a basic function in a cognitive radio [39, 61, 633]. Large random matrices are connected with spectral sensing through matrix functions [39, 61].

When the sizes of random matrices are large, concentration inequalities become inevitable. To account for this basic phenomenon, the monograph [40] develops the theory from the mathematical foundation to the algorithms.

Bibliographical Remarks

Communication is a critical ingredient for smart grid. It will require more discussion in future.

Early important references include [546, 551, 552, 634, 635] and the IEEE special issue in [636].

Section 15.3 draws material from [637]and [638]. We mainly follow the notation of [637].

Example 15.3.1 is taken from [639], [640] and [410] with some modification for the massive MIMO, to illustrate how the massive MIMO can be viewed to mimic the large CDMA systems.

In Section 15.4.1, we draw material from [213] and [286] to put in our context.

The results in Section 5.8 can be found in [139].

The paper [641] gives a first glimpse into, and opens up future work in, many unexploited research areas of applying wireless sensor networks (WSNs) in smart grid by providing an overview of the opportunities and challenges. In Section 10.9, we draw material from [641].

Distributed detection and estimation in wireless sensor networks is reviewed in [46]. In Section 16.1, we draw material from this paper. Section 16.3 draws from [642].

16

Big Data for Sensing

In [40], we treat a sensing system as a big data system and model the massive amount of data with the aid of large random matrices—see Fig. 1.6 for illustration. This chapter takes the same viewpoint and complements the above book.

16.1 Distributed Detection and Estimation

An emerging field is the use of wireless sensor networks (WSNs) as a support for smart grids. In such a case, a WSN is useful to: (i) monitor and predict energy production from renewable sources of energy such as wind or solar energy; (ii) monitor energy consumption; (iii) detect anomalies in the network.

Our problem can be stated as follows: given a large number of sensor nodes (say from tens to hundreds), how can we obtain functions of estimation and detection in a distributed manner? A sensor is defined in a very general sense. Examples include wireless sensors and cognitive radios. A cognitive radio deals with much more data [44].

16.1.1 Computing while Communicating

In a very general setting, taking a decision based on the data collected by the sensors can be interpreted as computing a function of these data. Let us denote by $x_i, i = 1, \ldots, N$, by the measurements collected by the i-th node of the network, and by $f(\mathbf{x}) = f\left(x_1, \ldots, x_N\right)$ the function to be computed.

To exploit the structure of the function $f(\mathbf{x}) = f\left(x_1, \ldots, x_N\right)$ to be computed, it is necessary to define some relevant structural properties. One important property is *divisibility*. Let $\mathcal{C}$ be a subset of $\{1, 2, \ldots, N\}$ and let $\pi := \left\{C_1, \ldots, C_s\right\}$ be a partition of $\mathcal{C}$. We denote by $\mathbf{x}_{C_i}$ the vector composed by the set of measurements collected by the nodes whose indices belong to C_i. A function $f(\mathbf{x}) = f\left(x_1, \ldots, x_N\right)$ is said to be *divisible*, for any $\mathcal{C} \subset \{1, 2, \ldots, N\}$ and any partition π, there exists a function $g^{(\pi)}$ such that

$$f\left(\mathbf{x}_C\right) = g^{(\pi)}\left(f\left(\mathbf{x}_{C_1}\right), f\left(\mathbf{x}_{C_2}\right), \ldots, f\left(\mathbf{x}_{C_s}\right)\right) \tag{16.1}$$

In words, (16.1) represents a sort of "divide and conquer" property: a function $f(\mathbf{x})$ is divisible if it is possible to split its computation into partial computations over subsets of data and then recombine the partial results to yield the desired outcome.

The idea of mingling computations and communications was proposed in [643]. An interesting link is established in [644] between the properties of the function $f(\mathbf{x})$ to

Smart Grid using Big Data Analytics: A Random Matrix Theory Approach, First Edition.
Robert C. Qiu and Paul Antonik.

be computed by the network and the topology of the communication network. Let $\mathcal{R}\left(f,N\right)$ be the range of $f\left(\mathbf{x}\right)$ and $\left|\mathcal{R}\left(f,N\right)\right|$ the cardinality of $\mathcal{R}\left(f,N\right)$. Under the following assumptions, we have

A.1 $f\left(\mathbf{x}\right)$ is divisible;
A.2 the network is connected;
A.3 the degree of each node is chosen as $d\left(N\right) \leqslant k_1 \log\left|\mathcal{R}\left(f,N\right)\right|$;

then, the rate for computing $f\left(\mathbf{x}\right)$ scales with N as

$$R\left(N\right) \geqslant \frac{c_1}{\log\left|\mathcal{R}\left(f,N\right)\right|} \tag{16.2}$$

Data uploading. Suppose it is necessary to convey all the data to the sink node. If each observed vector belongs to an alphabet $\mathcal{X}$ with cardinality $|\mathcal{X}|$, the cardinality of the whole data set is $\left|\mathcal{R}\left(f,N\right)\right| = |\mathcal{X}|^N$. So, $\log\left|\mathcal{R}\left(f,N\right)\right| = N\log|\mathcal{X}|$. From (16.2), the capacity of the network scales as $1/N$.

Decision based on the histogram of the measurements. Let us suppose now that the decision to be taken at the control node can be based on the histogram of the data collected by the nodes, with no information loss. In this case, the function $f\left(\mathbf{x}\right)$ is the histogram. It can be verified that the histogram is a divisible function. In this case the rate in (16.2) scales as $1/\log N$. If the decision can be based on the histogram of the data, rather than on each single measurement, adopting the right communication scheme, the rate per node behaves $1/\log N$, rather than $1/N$, with a rate gain $N/\log N$.

Symmetric functions. Let us consider the case where $f\left(\mathbf{x}\right)$ is a symmetric function. We recall that a function $f\left(\mathbf{x}\right)$ is symmetric if it is invariant to permutations of its arguments: $f\left(\mathbf{x}\right) = f\left(\mathbf{\Pi x}\right)$ for any permutation matrix $\mathbf{\Pi}$ and any argument vector $\mathbf{x}$. This property reflects the so called datacentric view. Examples of symmetric functions include the mean, median, maximum/minimum, and the histogram. The key property of symmetric functions is that it can be shown that they depend on the argument $\mathbf{x}$ only through the histogram of $\mathbf{x}$. Hence, the computation of symmetric functions is a particular case of the example examined before. Thus, the rate scales again as $1/\log N$.

16.1.2 Distributed Detection

Consider the hypothesis-testing problem

$$\mathcal{H}_0 : p\left(\mathbf{x}_1,\ldots,\mathbf{x}_N;\mathcal{H}_0\right)$$
$$\mathcal{H}_1 : p\left(\mathbf{x}_1,\ldots,\mathbf{x}_N;\mathcal{H}_1\right)$$

where $p\left(\mathbf{x}_1,\ldots,\mathbf{x}_N;\mathcal{H}_0\right)$ and $p\left(\mathbf{x}_1,\ldots,\mathbf{x}_N;\mathcal{H}_1\right)$ the joint probability density function of the whole set of observed data, under the hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively. The likelihood ratio test amounts to comparing the likelihood ratio (LR) with a threshold γ, and decide for $\mathcal{H}_1$, if the threshold is exceeded or for $\mathcal{H}_0$, otherwise. In formulas

$$\Lambda\left(\mathbf{x}\right) := \Lambda\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right) = \frac{p\left(\mathbf{x}_1,\ldots,\mathbf{x}_N;\mathcal{H}_1\right)}{p\left(\mathbf{x}_1,\ldots,\mathbf{x}_N;\mathcal{H}_0\right)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma \tag{16.3}$$

The LR test (LRT) is asymptotically optimal under a Bayes or a Neyman–Pearson criterion [423], in the sense that the length of random vector $\mathbf{x}$ goes infinite.

Let us now assume that the observations taken by different sensors are *statistically independent*, conditioned to each hypothesis. This is an assumption valid in many cases. Under such an assumption, the LR can be factorized as follows

$$\Lambda(\mathbf{x}) := \frac{\prod_{n=1}^{N} p\left(\mathbf{x}_n; \mathcal{H}_1\right)}{\prod_{n=1}^{N} p\left(\mathbf{x}_n; \mathcal{H}_0\right)} = \prod_{n=1}^{N} \Lambda_n\left(\mathbf{x}_n\right) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma \tag{16.4}$$

where

$$\Lambda_n\left(\mathbf{x}_n\right) = \frac{p\left(\mathbf{x}_n; \mathcal{H}_1\right)}{p\left(\mathbf{x}_n; \mathcal{H}_0\right)}$$

denotes the **local** LR at the n-th node. In this case, the global function $\Lambda(\mathbf{x})$ in (16.4) possesses a clear structure: it is factorizable in the product of the local LR functions. A factorizable function is also divisible.

The logarithm of the likelihood ratio can be written as

$$\log \Lambda\left(\mathbf{x}_1, \ldots, \mathbf{x}_N\right) = \sum_{i=1}^{N} \log \Lambda_i\left(\mathbf{x}_i\right) = \sum_{i=1}^{N} \log\left[\log p_{X_i}\left(\mathbf{x}_i; \mathcal{H}_1\right) - \log p_{X_i}\left(\mathbf{x}_i; \mathcal{H}_0\right)\right] \tag{16.5}$$

This formula shows that, in the conditionally independent case, running a consensus algorithm is sufficient to enable every node to compute the global LR. It is only required that every sensor initializes its own state with the local log-LR $\log \Lambda_i\left(\mathbf{x}_i\right)$ and then runs the consensus iterations. If the network is connected, every node will end up with the average value of the local LRs.

16.1.3 Distributed Estimation

Let us denote by $\boldsymbol{\theta} \in \mathbb{R}^N$ the parameter vector to be estimated. In some cases, there is no prior information about $\boldsymbol{\theta}$. In other cases, $\boldsymbol{\theta}$ is known to belong to a given set $\mathcal{C}$. In some applications, $\boldsymbol{\theta}$ may be the outcome of a random variable described by a known pdf $p_\Theta(\boldsymbol{\theta})$.

Let us denote by $\mathbf{x}_i$ the measurement vector collected node i and by $\mathbf{x} := \left[\mathbf{x}_1^T, \ldots, \mathbf{x}_N^T\right]$ the whole set of data collected by all the nodes. The estimation is obtained as the solution of the following optimization problem

$$\underset{\boldsymbol{\theta}}{\text{maximize}} \quad p_{X|\Theta}(\mathbf{x}|\boldsymbol{\theta})\, p_\Theta(\boldsymbol{\theta}) \tag{16.6}$$

where $p_\Theta(\boldsymbol{\theta})$ is the (known) prior pdf of the parameter vector and $p_{X/\Theta}(\mathbf{x}|\boldsymbol{\theta})$ is the pdf of $\mathbf{x}$ conditioned to $\boldsymbol{\theta}$. Let us consider the case where the pdf can be factorized as

$$p_{X;\Theta}(\mathbf{x}; \boldsymbol{\theta}) = g\left[\mathbf{T}(\mathbf{x}), \boldsymbol{\theta}\right] h(\mathbf{x}) \tag{16.7}$$

where $g(\cdot, \cdot)$ depends on $\mathbf{x}$ only through $\mathbf{T}(\mathbf{x})$ where $h(\cdot)$ does not depend on $\boldsymbol{\theta}$. The function $\mathbf{T}(\mathbf{x})$ is called a *sufficient statistic* for $\boldsymbol{\theta}$ [423].

A simple (yet common) example is given by the so called exponential family of pdf

$$p(\mathbf{x}; \boldsymbol{\theta}) = \exp\left[A(\boldsymbol{\theta})\, B(\mathbf{x}) + C(\mathbf{x}) + D(\boldsymbol{\theta})\right] \tag{16.8}$$

Examples of random variables described by this class include the Gaussian, Rayleigh, and exponential pdfs. Let us assume now that the observations $\mathbf{x}_i$ collected by different nodes are statistically independent and identically distributed (i.i.d.), according to (16.8). It is easy to check, simply applying the definition in (16.7), that a sufficient statistic in such a case is the scalar function:

$$T(\mathbf{x}) = \sum_{i=1}^{N} B\left(\mathbf{x}_i\right) \tag{16.9}$$

This structure suggests that a simple distributed way to enable every node in the network to estimate the vector $\boldsymbol{\theta}$ locally, without loss of optimality with respect to the centralized approach, is to run a *consensus algorithm*, where the initial state of every node is set equal to $B\left(\mathbf{x}_i\right)$. At convergence, if the network is connected, every node has a state equal to the consensus value: $T(\mathbf{x})/N$. This enables every node to implement the optimal estimation by simply interacting with its neighbors to achieve a consensus. The only necessary condition for this simple method to work properly is that the network is connected. This is indeed a very simple example illustrating how consensus can be a fundamental step in deriving an optimal estimation through a purely decentralized approach relying only upon the exchange of data among neighbors.

16.1.4 Consensus Algorithms

Consensus algorithms are fundamental to distributed algorithms including detection, estimation, and computing [645].

Given a set of measurements $x_i(0)$, for $i = 1, \ldots, N$, collected by the network nodes, the goal of a consensus algorithm is to minimize the disagreement among the nodes. This can be useful, for example, when the nodes are measuring some common variable and their measurement is affected by error. The scope of the interaction among the nodes is to reduce the effect of local errors on the final estimate. Consensus is one of fundamental tools to design distributed decision algorithms that satisfy a global optimality principle, as corroborated by many works on distributed optimization.

The proper way to describe the interactions among the network nodes is to introduce the graph model of the network.

Example 16.1.1 (the graph model for *N* sensors) Let us consider a network composed of N sensors. The flow of information across the sensing nodes implementing some form of distributed computation can be properly described by introducing a graph model whose vertices are the sensors, and there is an edge between two nodes if they exchange information with each other. Let us denote the graph by $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ where $\mathcal{G}$ denotes the set of N vertices (nodes) v_i and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges $e_{ij}\left(v_i, v_j\right)$.

The most powerful tool to grasp the properties of a graph is *algebraic graph theory* [646], which is based on the description of the graph through appropriate matrices. Let $\mathbf{A} \in \mathbb{R}^{N \times N}$ be the *adjacency matrix* of the graph $\mathcal{G}$, whose elements a_{ij} represent the weights associated to each edge with $a_{ij} > 0$ if $e_{ij} \in \mathcal{E}$ and $a_{ij} = 0$ otherwise. According to this notation and assuming no self-loops, i.e., $a_{ii} = 0$, $\forall i = 1, \ldots, N$, the out-degree of node v_i is defined as $\deg_{out}\left(v_i\right) = \sum_{j=1}^{N} a_{ji}$. Similarly, the in-degree of node v_i is $\deg_{in}$

$\left(v_i\right) = \sum_{j=1}^{N} a_{ji}$. The *degree matrix* $\mathbf{D}$ is defined as the diagonal matrix whose i-th diagonal entry is $d_{ii} = \deg\left(v_i\right)$. Let $\mathcal{N}_i$ denote the set of neighbors of node i, so that $\left|\mathcal{N}_i\right| = \deg_{in}\left(v_i\right)$, where by $|\cdot|$ we denote the cardinality of the set. The *Laplace matrix* of $\mathbf{L} \in \mathbb{R}^{N\times N}$ is defined as

$$\mathbf{L} := \mathbf{D} - \mathbf{A}$$

Some properties of the Laplacian will be used in the distributed algorithms to be presented later on, and then it is useful to recall them.

Properties of the Laplacian matrix $\mathbf{L}$ include:

P.1: $\mathbf{L}$ has, by construction, a null eigenvalue with associated eigenvector the vector $\mathbf{1}$ composed by all ones. This property can be easily checked verifying that $\mathbf{L}1 = 0$ since by construction, $\sum_{j=1}^{N} a_{ij} = d_{ii}$.

P.2: The multiplicity of the null eigenvalue is equal to the number of connected components of the graph. Hence, the null eigenvalue is simple (it has multiplicity one) if and only if the graph is connected.

P.3: If we associate a state variable x_i to each node of the graph, if the graph is undirected, the disagreement between the values assumed by the variables is a quadratic form built on the Laplacian [646]:

$$J(\mathbf{x}) = \frac{1}{4}\sum_{i=1}^{N}\sum_{j\in\mathcal{N}_i} a_{ij}\left(x_i - x_j\right)^2 = \frac{1}{2}\mathbf{x}^T\mathbf{L}\mathbf{x} \tag{16.10}$$

where $\mathbf{x} := \left[\mathbf{x}_1^T, \ldots, \mathbf{x}_N^T\right]$ denote the network state vector and $\mathcal{N}_i$ indicates the set of neighbors of node i. □

Example 16.1.2 (average values) The nodes are measuring a temperature and the goal is to find the average temperature. In this case, reaching a consensus over the average temperature can be seen as the minimization of the disagreement, as defined in (16.10), between the states $x_i(0)$ associated with the nodes. The minimization of the disagreement can be obtained by using a simple gradient-descent algorithm. More specifically, using a continuous-time system, the minimum of (16.10) can be achieved by running the following dynamic system

$$\frac{d\mathbf{x}(t)}{dt} = -\mathbf{L}\mathbf{x}(t) \tag{16.11}$$

initialized with $\mathbf{x}(0) = \mathbf{x}_0$, where $\mathbf{x}_0$ is the vector containing all the initial measurements collected by the network nodes. This means that the state of each node evolves in time according to the first order differential equation

$$\frac{dx_i(t)}{dt} = \sum_{j\in\mathcal{N}_i} a_{ij}\left(x_i - x_j\right) \tag{16.12}$$

Hence, every node updates its own state only by interacting with its neighbors.

The solution of (16.11) is given by

$$\mathbf{x}(t) = \exp(-\mathbf{L}t)\,\mathbf{x}(0) \tag{16.13}$$

The convergence of (16.13) is guaranteed because all the eigenvalues of $\mathbf{L}$ are non-negative, by construction. If the graph is connected, using **P.2**, the eigenvalue zero has multiplicity one. Furthermore, the eigenvector associated to the zero eigenvalue is the vector 1. Hence, the system (16.11) converges to the consensus state:

$$\lim_{t\to\infty} \mathbf{x}(t) = \frac{1}{N}\mathbf{1}\mathbf{1}^T\mathbf{x}(0)$$

This means that every node converges to the average value of the measurements collected by the whole network:

$$\lim_{t\to\infty} \mathbf{x}(t) = \frac{1}{N}\sum_{i=1}^{N} x_i(0) = x^*$$

Alternatively, the minimization of (16.13) can be achieved in discrete-time through the following iterative algorithm

$$\mathbf{x}[k+1] = \mathbf{x}[k] - \varepsilon\mathbf{L}\mathbf{x}[k] := \mathbf{W}\mathbf{x}[k] \tag{16.14}$$

where $\mathbf{W} = \mathbf{I} - \varepsilon\mathbf{L}$ is the transition matrix. In this case, the discrete time equation is initialized with the measurements taken by the sensor nodes at time 0, i.e., $\mathbf{x}[0] := \mathbf{x}_0$. The convergence is guaranteed. □

16.1.5 Random Geometric Graph with Euclidean Random Matrix (ERM)

A random graph is obtained by distributing N points randomly over the d-dimensional space $\mathbb{R}^d$ and connecting the nodes according to a given rule. The graph topology is captured by the adjacency matrix $\mathbf{A}$, which, in this case, is a random matrix. An important class of Random Matrices, is the so called Euclidean random matrix (ERM) class. See also Section 6.14 and Section 16.2. Given a set of N points located at positions $\mathbf{x}_i$, $i = 1, \ldots, N$, an $N \times N$ adjacency matrix $\mathbf{A}$ is an ERM if its generic (i, j) entry depends only on the difference $\mathbf{x}_i - \mathbf{x}_j$:

$$a_{ij} = F\left(\mathbf{x}_i - \mathbf{x}_j\right)$$

where F is a measurable mapping from $\mathbb{R}^d$ to $\mathbb{R}$. An important subclass of ERM is given by the adjacency matrices of the so called random geometric graphs (RGG). In such a case, the entries a_{ij} of the adjacency matrix are either zero or one depending only on the distance between nodes i and j:

$$a_{ij} = F\left(\mathbf{x}_i - \mathbf{x}_j\right) = \begin{cases} 1 & \text{if } \left\|\mathbf{x}_i - \mathbf{x}_j\right\| \leqslant r \\ 0 & \text{otherwise} \end{cases} \tag{16.15}$$

where r is the coverage radius.

Example 16.1.3 (concentration of spectral measure for a random geometric graph) Assuming that the RGG $G(N, r)$ is connected with high probability, [647] derived an analytical expression for the algebraic connectivity of the graph: the second eigenvalue of the symmetric Laplacian, $\mathbf{L} = D - A$, where $\mathbf{D}$ is the degree matrix and $\mathbf{A}$ is the adjacency matrix.

The analytical tool they used is the concentration of spectral measure [40]. In [648,649], it was shown that the eigenvalues of the adjacency matrix tend to be

concentrated, as the number of nodes tend to infinity. In [649], it was shown that the eigenvalues of the normalized adjacency matrix $\mathbf{A}_N = \mathbf{A}/N$ of an RGG $G(N, r)$, composed of points uniformly distributed over a unitary two-dimensional torus, tend to the Fourier series

$$\hat{F}(\mathbf{z}) = \int_{\Omega_r} \exp\left(-j2\pi \mathbf{z}^T \mathbf{x}\right) d\mathbf{x}$$

coefficients of the function F defined in (16.15), almost surely, for all $\mathbf{z} = \left[z_1, z_2\right] \in \mathbb{Z}^2$, where $\Omega_r = \left\{\mathbf{x} = \left[x_1, x_2\right]^T \in \mathbb{R}^2 : \|\mathbf{x}\| \leqslant r\right\}$. □

16.2 Euclidean Random Matrix

The Euclidean random matrix is introduced in [363, 649, 650]. See also the outstanding work of [321, 358, 360] and the PhD dissertation [360]. An $n \times n$ Euclidean random matrix $\mathbf{A}_n = \left(A_{ij}\right)_{n\times n}$ is defined in [364, 651, 652] with the help of some function g of n points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ that are randomly distributed to $g\left(\mathbf{x}_i, \mathbf{x}_j\right)$. The elements A_{ij} are equal to $g\left(\mathbf{x}_i, \mathbf{x}_j\right)$.

Euclidean random matrices play an important role in description of many physical models including the electronic levels in amorphous systems, very diluted impurities, and the spectrum of vibrations in glasses. Here we want to point out the connection between the Euclidean random matrix and the massive MIMO (or wireless sensor network).

Here we point out a special class of Euclidean random matrices such that [651]

$$\mathbf{M}_n = \left(f_n\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right)_{n\times n} \tag{16.16}$$

where $f_n(x)$ is a real function defined on $[0, \infty)$, and $||\cdot||$ is the Euclidean distance with $\|\mathbf{x}\| = \sqrt{x_1^2 + \cdots x_N^2}$ for any vector $\mathbf{x} = \left(x_1, .., x_N\right) \in \mathbb{R}^N$.

Some literature already exists on the Euclidean matrix model (1.1) for different functions of $f_n(x)$. For example, [650] first considered the Gaussian Euclidean random matrix corresponding to $f_n(x) = (2\pi)^{-3/2} \exp(-x/2)$ in (16.16). And taking $f_n(x) = \sqrt{x}$ for all $n \geq 2$, (16.16) is reduced to

$$\mathbf{B}_n = \left(\left\|\mathbf{x}_i - \mathbf{x}_j\right\|\right)_{n\times n} \tag{16.17}$$

which is referred as the Euclidean distance matrix. There are also cardinal sine Euclidean random matrices and cosine Euclidean random matrices with

$$f_n(x) = \sin\left(k_0\sqrt{x}\right)/\left(k_0\sqrt{x}\right) \text{ and } f_n(x) = \left(1 - \delta_{ij}\right)\cos\left(k_0\sqrt{x}\right)/\left(k_0\sqrt{x}\right)$$

where k_0 is a constant, and δ_{ij} is the Kronecker symbol. One can study the exponential Euclidean random matrices with $f_n(x) = \exp\left(-\sqrt{x}/\xi\right)$, where ξ is the location length.

Very recently, Jiang [364] investigated Euclidean random matrices with random vectors generated from geometrical shapes G and obtained some important and interesting results. Specifically:

1) When the dimension N of the geometry G is fixed and the number of sample points $n \to \infty$, Jiang [364] showed the empirical distribution of the eigenvaluesof $\mathbf{M}_n$

converges weakly to δ_0 for a big class functions of $f_n(x)$. Further, the conclusion holds regardless of the shape of $\mathbf{G}$.

2) When $N = N(n)$ becomes large as n increases, some simulations made in [364] showed that the behavior of the empirical spectral distribution of $\mathbf{M}_n$ depends on the topology of $\mathbf{G}$. Further, when $\mathbf{M}_n$ is generated from l_p unit ball $B_{N,p}$ or sphere $S_{N,p}$ with $p \geq 1$ and both N and n go to infinity proportionally, Jiang [364] derived the explicit nice expression for the limiting spectral distribution of scaled $\mathbf{M}_n$. It is in the form of $a + bV$ where a, b are constants and V has the famous Marchenko– Pastur distribution. And the condition on $f_n(x)$ is that $f_n(x)$ is locally twice differential at an explicit value.

Here $B_{N,p}$ or sphere $S_{N,p}$ are defined as

$$B_{N,p} = \left\{\mathbf{y} \in \mathbb{R}^N; \|\mathbf{y}\|_p \leqslant 1\right\} \text{ and } S_{N,p} = \left\{\mathbf{y} \in \mathbb{R}^N; \|\mathbf{y}\|_p = 1\right\}$$

16.3 Decentralized Computing

Due to the distributed nature of the 5G network, decentralized computing is critical. Very often, algorithms only depend on *eigenvalues* of these random matrices. We make this assumption to simplify the algorithms required for distributed computing.

Computing is critical to real-time applications such as detection and estimation. Often the eigenvalues of large random matrices need be computed in real time. As far as we are aware, all previous works (except a few papers) assume a centralized architecture, in which a fusion center gathers the signal samples received by all sensors, processes the data, and forwards the decision back to all nodes. Such an architecture suffers from scalability issues and is not suitable for large-scale, multihop networks, where the fusion center may be many hops away from peripheral nodes.

For this reason, we seek a decentralized implementation of eigenvalue-based applications, such that the computational effort is distributed across multiple sensor nodes communicating iteratively with neighbors (i.e., gossip algorithms) and the algorithm's statistics are computed locally by each node. We propose two solutions based on iterative eigenvalue algorithms—the power method and the Lanczos algorithm. In our large-scale network testbed, these algorithms can be implemented in a distributed fashion by applying distributed average consensus algorithms. The only relevant work is listed in [630–632, 642].

Consider a wireless network consisting of K sensor nodes. During a given time interview (sensing period) each node collects N complex signal samples. Our typical values for N and K vary from tens to hundreds and to even thousands. The global received sample matrix is denoted by

$$\mathbf{Y} = \left[\mathbf{y}_1, \ldots, \mathbf{y}_N\right] = \begin{bmatrix} \mathbf{y}[1]^T \\ \vdots \\ \mathbf{y}[K]^T \end{bmatrix} \in \mathbb{C}^{K \times N} \tag{16.18}$$

where symbols $\mathbf{y}_i \in \mathbb{C}^{K \times 1}, i = 1, \ldots, N$ and $\mathbf{y}\left[k\right] \in \mathbb{C}^{N \times 1}, k = 1, \ldots, K$ are used to denote, respectively, the columns and (transpose) rows of $\mathbf{Y}$. Physically, column $\mathbf{y}_i, i = 1, \ldots, N$

contains the samples received by all nodes at time $t = iT_s$, (where $1/T_s$ is the sampling rate), whereas row $\mathbf{y}\left[k\right]^T$ contains all samples available at node k at the end of the sensing period. We then define the sample covariance matrix as

$$\mathbf{R} \triangleq \frac{1}{N}\mathbf{Y}\mathbf{Y}^H \tag{16.19}$$

Let $\lambda_1 \geqslant \cdots \geqslant \lambda_K \geqslant 0$ be the eigenvalues of $\mathbf{R}$, without loss of generality sorted in decreasing order, and $\mathbf{u}_1, \ldots, \mathbf{u}_k$ the corresponding eigenvectors. The problem can be stated as follows:

Problem statement for decentralized computing. How can a network compute (or estimate) one or more of the above eigenvalues without a fusion center that collects all samples (data matrix) $\mathbf{Y}$, and *without explicitly constructing* the sample covariance matrix $\mathbf{R}$?

Penna and Stanczak [642] derive and analyze two general-purpose algorithms—referred to as the decentralized power method (DPM) and the decentralized Lanczos algorithm (DLA)—for distributed computation of one (the largest) or multiple eigenvalues of a sample covariance matrix over a wireless network ($K = 40$ nodes and $N = 10$ samples per node were assumed in [642]). Given the increasing popularity of dense, large-scale wireless sensor networks, applications of eigenvalue-based inference techniques in distributed settings are of great interest. They seek a decentralized method to compute the eigenvalues of sample covariance matrices over a wireless network, such that the computational effort is distributed across multiple nodes and the many-to-one communication protocol is replaced by a more scalable neighbor-to-neighbor protocol. Eigenvalue-based hypothesis tests can be implemented in a decentralized setting by using the proposed algorithms—the DPM and the DLA. Such decentralized signal detection techniques enable sensor nodes to compute global test statistics locally, thereby performing hypothesis tests without relying on a fusion center.

The popular cooperative energy detector, the test based on the (possibly weighted) sum of the received signal energies at different sensors, also admits a natural decentralized implementation via average consensus algorithms [653]. This problem was investigated in [654]. Decentralized energy detection is computationally simpler than eigenvalue-based techniques, but clearly inherits the well-known shortcomings of energy detection (suboptimality in multisensor settings and sensitivity to noise uncertainty).

Possible statistics include:

- Roy's largest root test: the largest eigenvalue normalized by the noise variance (optimal under the Neyman–Pearson criterion: λ_1/σ_v^2;
- the generalized likelihood ratio test (GLRT) statistic: $\lambda_1 / \left(\sum_{i=2}^{K} \lambda_i \right)$;
- the sphericity test statistic: $\left(\prod_{i=1}^{K} \lambda_i \right) / \left(\frac{1}{K} \sum_{i=1}^{K} \lambda_i \right)$. Taking the logarithm, we obtain

$$\sum_{i=1}^{K} \log \lambda_i - \log \left(\frac{1}{K} \sum_{i=1}^{K} \lambda_i \right) = \mathrm{Tr}\left[\log\left(\mathbf{R}\right)\right] - \log\left[\frac{1}{K}\,\mathrm{Tr}\left(\mathbf{R}\right)\right]$$

- John's test: $\sum_{i=1}^{K} \lambda_i^2 / \left(\sum_{i=1}^{K} \lambda_i \right)^2$. Or $\mathrm{Tr}\, \mathbf{R}^2 / (\mathrm{Tr}\, (\mathbf{R}))^2$.

In the sphericity test statistic, we end up with the trace function $\mathrm{Tr} f\left(\mathbf{X}\mathbf{X}^H\right)$ where $f : \mathbb{R}^+ \to \mathbb{R}^+$ is some continuous function. The concentration of spectral measure phenomenon occurs for this type of trace matrix functions. See Qiu and Wicks [40] for details.

When the number of sensor nodes K is large, few samples N per node are needed to achieve high detection performance, thus reducing the size of messages exchanged between neighboring nodes.

For distributed estimation in a wireless sensor network, multiple spatially distributed sensors collaborate to estimate the system state of interest, without the support of a central fusion center due to physical constraints such as large system size and limited communications infrastructure. Specifically, each sensor makes local partial observations and communicates with its neighbors to exchange certain information, in order to enable this collaboration. Due to its scalability for large systems, and robustness to sensor failures, distributed estimation techniques find promising and wide applications including in battlefield surveillance, environment sensing, and power-grid monitoring. Especially in the era of *big data and large systems*, which usually require overwhelming computation if implemented in a centralized way, distributed schemes become critical because they can decompose the computational burden into local parallel procedures. A principal challenge in distributed sensing, and in distributed estimation in particular, is to design the distributed algorithm to achieve reliable and mutually agreeable estimation results across all sensors, without the help of a central fusion center. See [655–657] for the above concerns. See [653, 658] for recent surveys.

Appendix A: Some Basic Results on Free Probability

A.1 Non-Commutative Probability Spaces

Let $\mathcal{A}$ be an algebra of operators that act on a Hilbert space. We will assume that $\mathcal{A}$ contains the identity operator (such algebras are called unital) and that it is closed under the operation of taking adjoints, that is, if $\mathbf{X} \in \mathcal{A}$, then $\mathbf{X}^* \in \mathcal{A}$.

It is often convenient to assume further that $\mathcal{A}$ is closed either with respect to uniform operator norm ($\|\mathbf{X}\| = \sup_{\|\mathbf{v}\|=1} \|\mathbf{X}\mathbf{v}\|$), or with respect to weak topology ($\mathbf{X}_i \to \mathbf{X}$ if and only if $\langle \mathbf{u}, \mathbf{X}_i \mathbf{v} \rangle \to \langle \mathbf{u}, \mathbf{X}\mathbf{v} \rangle$ for all vectors $\mathbf{u}$ and $\mathbf{v}$).

In the first case, the algebra is called a C^*-algebra, and in the second case, it is called a W^*-algebra or a von Neumann algebra.

A *state* on the algebra $\mathcal{A}$ is a **linear** functional $E : \mathcal{A} \to \mathbb{C}$, which has the following positivity property: for all operators $\mathbf{X}$

$$E(\mathbf{X}^*\mathbf{X}) \geqslant 0$$

A typical example of a state is $E(\mathbf{X}) = \langle \mathbf{u}, \mathbf{X}\mathbf{u} \rangle$ where $\mathbf{u}$ is a unit vector. Typically, a state is denoted by letters φ or τ but we will use letter E to emphasize the parallel with expectation functional $\mathbb{E}$ in the classical probability theory.

The name "state" is due to the relation of operator algebras to quantum mechanics. In this Appendix we will use the words "state" and "expectation" interchangeably. States may have some additional properties. If $E(\mathbf{A}^*\mathbf{A}) = 0$ implies that $\mathbf{A} = 0$, then the state is called *faithful*. If $\mathbf{X}_i \to \mathbf{X}$ weakly implies that $E\left(\mathbf{X}_i\right) \to E(\mathbf{X})$ then the state is called *normal*. If $E(\mathbf{XY}) = E(\mathbf{YX})$, then the state is called *tracial*, or simply *trace*.

Definition A.1.1 A noncommutative probability space $(\mathcal{A}, E)$ is a pair of a unital C^*-operator algebra $\mathcal{A}$ and a state E with additional property $E(\mathbf{I}) = 1$.

If the state E is tracial, then we call $(\mathcal{A}, E)$ a *tracial* noncommutative probability space. If $\mathcal{A}$ is a von Neumann algebra and E is normal, then we call $(\mathcal{A}, E)$ a *W^*-probability space*.

Here are several examples.

- **A classical probability space.**
- **The algebra of N-by-N matrices.** In this case, we can use the normalized trace[1] as the expectation:

1 We use Tr $(\cdot)$ to denote the unnormalized trace.

Smart Grid using Big Data Analytics: A Random Matrix Theory Approach, First Edition.
Robert C. Qiu and Paul Antonik.

$$E\mathbf{X} = \mathrm{tr}\,(\mathbf{X}) := \frac{1}{N}\sum_{i=1}^{N} X_{ii}$$

- **The algebra of random N-by-N matrices.** The joint probability distribution of the entries of these matrices is such that all joint moments of these entries are finite. We define the functional E as the expectation of the trace:

  $$E\mathbf{X} = \langle \mathrm{tr}\,(\mathbf{X}) \rangle$$

  Here we used a convenient notation from the physical literature: $\langle Z \rangle$ denotes the average of Z over the statistical ensemble, that is, the expectation of the random variable Z. We often use $\mathbb{E}\,(Z)$ to represent $\langle Z \rangle$, too.

A.2 Distributions

Suppose that $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are elements of a noncommutative probability space $(\mathcal{A}, E)$. We will call them random variables. Their *distribution* is the **linear** map from the algebra of polynomials in non-commuting variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ to $\mathbb{C}$

$$f\left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\right) \to E\left[f\left(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\right)\right]$$

The $*$-distribution is a similar map for polynomials in noncommutative variables $\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{y}_1, \ldots, \mathbf{y}_n$, which is given by the formula:

$$f\left(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{y}_1, \ldots, \mathbf{y}_n\right) \to E\left[f\left(\mathbf{X}_1, \ldots, \mathbf{X}_n, \mathbf{X}_1^*, \ldots, \mathbf{X}_n^*\right)\right]$$

In other words, the distribution of a family of random variables is a collection of their joint moments.

Proposition A.2.1 Suppose that $\mathbf{X}$ is a bounded self-adjoint element of a W^*-probability space $(\mathcal{A}, E)$. Then there exists a probability measure μ on $\mathbb{R}$ such that

$$E\left(\mathbf{X}^k\right) = \int_{\mathbb{R}} x^k \mu\,(dx).$$

For an $N \times N$ random matrix $\mathbf{Z}$, we have

$$E\left(\mathbf{X}^k\right) = \frac{1}{N}\mathbb{E}\left[\mathrm{Tr}\left(\mathbf{Z}^k\right)\right] = \int_{\mathbb{R}} x^k \mu_N\,(dx)$$

Examples:

- If $\mathbf{X}$ is a Hermitian matrix, then

  $$\mu = \frac{1}{N}\sum_{i=1}^{N} \delta_{\lambda_i}$$

  where λ_i are eigenvalues of $\mathbf{X}$, counted with multiplicity.
- If $\mathbf{X}$ is a Hermitian random matrix consider as an element of noncommutative probability space, the spectral probability distribution of $\mathbf{X}$ is

  $$\mu = \frac{1}{N}\mathbb{E}\left(\sum_{i=1}^{N} \delta_{\lambda_i}\right)$$

A.3 Asymptotic Freeness of Large Random Matrices

For two random matrices, the statistical independence is replaced with asymptotic freeness in free probability theory.

Theorem A.3.1 Let $\mathbf{A}_N$ and $\mathbf{B}_N$ be $N \times N$ Hermitian matrices that converge in distribution to the pair $\{a, b\}$. Let $\mathbf{U}_N$ be a sequence of $N \times N$ independent random unitary matrices that have the Haar distribution on the unitary group $\mathcal{U}(N)$. Then $\mathbf{A}_N$ and $\mathbf{U}_N \mathbf{B}_N \mathbf{U}_N^H$ converge in distribution to random variables a and $\tilde{b}$, where $\tilde{b}$ has the same distribution as b, and a and $\tilde{b}$ are free.

A.4 Limit Theorems

The following theorem is an analog of the central limit theorem for sums of i.i.d. random variables.

Theorem A.4.1 (limit theorem for sums of free random variables) Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a sequence of identically distributed bounded self-adjoint (matrix-valued) random variables. Assume that $\mathrm{Tr}\left(\mathbf{X}_i\right) = 0$, $\mathrm{Tr}\left(\mathbf{X}_i^2\right) = 1$, and that $\mathbf{X}_i$ are free. Define $\mathbf{S}_n = \mathbf{X}_1 + \cdots + \mathbf{X}_n$. Then the sequence $\mathbf{S}_n / \sqrt{n}$ converges in distribution to the standard semicircle random variable.

Let us now discuss a free analog of another theorem from classical probability theory, which is sometimes called the law of small numbers. In the classical case, this law says that counts of rare events are distributed by the Poisson law.

Let us define the free analog of the Poisson law. Let μ be a distribution with the density

$$p(x) = \frac{1}{2\pi x} \sqrt{4x - (1 - c + x)^2} \text{ if } x \in \left[\left(1 - \sqrt{c}\right)^2, \left(1 + \sqrt{c}\right)^2 \right]$$

where c is is a positive parameter. If x is outside of this interval, then the density is zero. In addition, if $c < 1$, then the distribution has an atom at 0 with weight $(1 - c)$.

This distribution is called the *free Poisson distribution* with parameter c. It is also known as the Marchenko–Pastur distribution because it was discovered in [219].

Theorem A.4.2 Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be self-adjoint random variables with the Bernoulli distribution

$$\mu = \left(1 - \frac{c}{n}\right) \delta_0 + \frac{c}{n} \delta_1$$

Assume that $\mathbf{X}_i$ are free and define

$$\mathbf{S}_n = \mathbf{X}_1 + \cdots + \mathbf{X}_n$$

Then, $\mathbf{S}_n / \sqrt{n}$ converges in distribution to the free Poisson distribution with the parameter c.

A.5 *R*-diagonal Random Variables

Generalizations of Haar unitaries based on this property are called *R*-diagonal random variables. This generalization seems to be the simplest class of non-Hermitian operators which that be handled using the methods of free probability.

Theorem A.5.1 Suppose that $\mathbf{U}$ is a Haar unitary, $\mathbf{H}$ is an arbitrary operator, and $\mathbf{U}$ and $\mathbf{H}$ are free. Then the variable $\mathbf{X} = UH$ is *R*-diagonal.

Theorem A.5.2 Suppose that $\mathbf{X}$ is an *R*-diagonal element in a tracial noncommutative probability space. Then it can be represented in distribution by a product $\mathbf{U}H$ where $\mathbf{U}$ is a Haar unitary and $\mathbf{H}$ is a positive operator that has the same distribution as $\sqrt{\mathbf{X}^H\mathbf{X}}$.

The above theorems say that the sum and the product of two free *R*-diagonal elements is *R*-diagonal. What is more surprising is that powers of an *R*-diagonal element are *R*-diagonal. We write $\mathbf{A} \cong \mathbf{B}$ if two random matrices $\mathbf{A}$ and $\mathbf{B}$ are equivalent, meaning that they have the same $*$-distributions.

Theorem A.5.3 Let $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ be free *R*-diagonal random variables in a C^*-probability space $\mathcal{A}_1$ and $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_n$ be self-adjoint random variables in a probability space $\mathcal{A}_2$. Assume that $\mathbf{A}_i$ has the same probability distribution as $|\mathbf{a}_i| := \sqrt{\mathbf{a}_i^*\mathbf{a}_i}$ for every i. Let $\mathbf{U}$ be a Haar unitary in $\mathcal{A}_2$, which is free from $\{\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_n\}$. Let $\mathbf{\Pi} = \mathbf{a}_n \cdots \mathbf{a}_1$ and $\mathbf{X} = \mathbf{U}\mathbf{A}_n \cdots \mathbf{U}\mathbf{A}_1$. Then $\mathbf{\Pi} \cong \mathbf{X}$.

In other words, if we multiply variables $\mathbf{A}_i$ by the Haar-distributed rotation $\mathbf{U}$, then we will lose all dependencies and the distribution of the product of these variables will be the same as if they all were *R*-diagonal and free. The surprising fact is that $\mathbf{U}$ is the same for all $\mathbf{A}_i$.

An important particular case is when all $\mathbf{a}_i$ are identically distributed.

Proposition A.5.4 Suppose that $\mathbf{X}$ is *R*-diagonal. Then $\mathbf{X}^n$ is *R*-diagonal for every integer $n \geq 1$.

Proof: Let $\mathbf{X} \cong \mathbf{X}_i$ and $\mathbf{X}_i$ be free. Then $\mathbf{X}^n \cong \mathbf{X}_n \cdots \mathbf{X}_1$ and the product $\mathbf{X}_n \cdots \mathbf{X}_1$ is *R*-diagonal as a product of free *R*-diagonal elements. □

A.6 Brown Measure of *R*-diagonal Random Variables

We define a generalization of the eigenvalue distribution for infinite-dimensional non-normal operators. It is defined only for operators in von Neumann algebras and uses the fact that in these algebras one can define an analog of the determinant, which is called the Fuglede–Kadison determinant.

Definition A.6.1 Let $\mathbf{X}$ be a bounded random variable in a tracial W^*-probability space $(\mathcal{A}, E)$. Then the Fuglede–Kadison determinant of $\mathbf{X}$ is defined as

$$\det \mathbf{X} := \exp\left[\frac{1}{2} E \log\left(\mathbf{X}^*\mathbf{X}\right)\right]$$

Consider the algebra of $N \times N$ matrices with $E(\mathbf{X}) = \frac{1}{N} \operatorname{Tr}(\mathbf{X})$. Then, we can write the Fuglede–Kadison determinant as $\det \mathbf{X} = \left(\prod_{i=1}^{N} s_i \right)^{1/N}$, where s_i are the singular values of the matrix $\mathbf{X}^* \mathbf{X}$. By using results from linear algebra, we have

$$\det \mathbf{X} = (\operatorname{Det}(\mathbf{X}^* \mathbf{X}))^{\frac{1}{2N}} = |\operatorname{Det}(\mathbf{X})|^{1/N}$$

where $\operatorname{Det}(\mathbf{X})$ is the usual determinant. It follows that

$$\log \det(\mathbf{X} - \lambda \mathbf{I}) = \frac{1}{N} \sum_{i=1}^{N} \log |\lambda_i - \lambda|$$

Here λ_i are eigenvalues of $\mathbf{X}$ taken with multiplicities that are equal to the number of times that λ_i is repeated on the diagonal of the Jordan form of $\mathbf{X}$.

We can think about function $\log \det(\mathbf{X} - \lambda \mathbf{I})$ as a suitable generalization of the logarithm of the modulus of characteristic polynomial.

Definition A.6.2 The L-function of random variable $\mathbf{X}$ is defined as $L_{\mathbf{X}}(\lambda) := \log \det(\mathbf{X} - \lambda \mathbf{I}) = E \log |\mathbf{X} - \lambda \mathbf{I}|$, where $|\mathbf{X} - \lambda \mathbf{I}| = \left[(\mathbf{X} - \lambda \mathbf{I})^* (\mathbf{X} - \lambda \mathbf{I}) \right]^{1/2}$.

Definition A.6.3 Let $\mathbf{X}$ be a bounded random variable in a tracial W^*-probability space $(\mathcal{A}, E)$. Then its Brown measure is a measure $\mu_{\mathbf{X}}$ on the complex plane $\mathbb{C}$ defined by the following equation. Let $\lambda = x + iy$. Then

$$\mu_{\mathbf{X}} = \frac{1}{2\pi} \Delta L_{\mathbf{X}}(\lambda) \, dxdy$$

where $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ is the Laplace operator, and the equality holds in the sense of (Schwarz) distributions.

For an R-diagonal operator, the Brown measure is invariant with respect to rotations of the complex plane around the origin and we can write it as a product of the radial and polar part. Let us list (without proof) some of the properties of the Brown measure:

- It is a unique measure such that $L_{\mathbf{X}}(\lambda) = \int_{\mathbb{C}} \log |z - \lambda| d\mu_{\mathbf{X}}(z)$;
- For every integer $k \geq 0$, we have $E\left(\mathbf{X}^k\right) = \int_{\mathbb{C}} z^n d\mu_{\mathbf{X}}(z)$,
- The Brown measure of a normal operator $\mathbf{X}$ coincides with its spectral probability distribution.

Theorem A.6.4 Let $\mathbf{X}$ and $\mathbf{Y}$ be two bounded random variables in a tracial W^*-probability space $(\mathcal{A}, E)$. Then (i) $\det(\mathbf{XY}) = \det \mathbf{X} \det \mathbf{Y}$, and (ii) $\det\left(e^{\mathbf{X}}\right) = \left|e^{E(\mathbf{X})}\right| = \exp(\operatorname{Re} E(\mathbf{X}))$.

If $\mathbf{H}$ is a self-adjoint random variable, and $\mathbf{A}$ is an arbitrary bounded variable, then

$$\det\left[\exp(\mathbf{A}^*) \exp(\mathbf{H}) \exp(\mathbf{A})\right] = \exp(E(\mathbf{A}^* + \mathbf{A})) \det\left[\exp(\mathbf{H})\right]$$

Appendix B: Matrix-Valued Random Variables

At the heart of modeling big data is the methodology of using large random matrices as the basic building blocks. Then we study the distribution of the matrix-valued random variables.

B.1 Random Vectors and Random Matrices

Multivariate analysis deals with issues related to the observations of correlated random variables. We denote a set of p random variables $X_1, \ldots, X_p$ by a vector

$$\mathbf{X} = \left(X_1, \ldots, X_p\right)^T$$

which is called a random vector. The mean or expectation of $\mathbf{X}$ is defined to be the vector of expectations:

$$\mathbb{E}\left(\mathbf{X}\right) = \begin{pmatrix} \mathbb{E}\left(X_1\right) \\ \vdots \\ \mathbb{E}\left(X_p\right) \end{pmatrix}$$

A typical set of multivariate random samples, $\left\{\mathbf{X}_1, \ldots, \mathbf{X}_n\right\}$, arises from taking measurements on a $p \times 1$ random vectors $\mathbf{X}$ for each of n objects or people. It is convenient to express these observations in matrix form

$$\mathbf{X} = \left(\mathbf{X}_1, \ldots, \mathbf{X}_n\right)^T = \begin{pmatrix} \mathbf{X}_1^T \\ \mathbf{X}_2^T \\ \vdots \\ \mathbf{X}_n^T \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}_{n \times p}$$

Let $\mathbf{X}$ be a matrix of random variables, which we call a *random matrix*. Here the rows of $\mathbf{X}$ may or may not be random observations of $\mathbf{X}$. More generally, the expectation of a random matrix $\mathbf{X} = \left(X_{ij}\right)$ is defined by the matrix whose (i,j)-th element is $\mathbb{E}\left(X_{ij}\right)$: namely, $\mathbb{E}\left(\mathbf{X}\right) = \left(\mathbb{E}\left(X_{ij}\right)\right)$.

To model the large datasets, we often require n and p to be large but finite. In practice, both n and p are comparable, or

$$n \to \infty, p \to \infty \text{ but } \frac{p}{n} \to c \in [0, \infty)$$

We call these *large-dimensional random matrices* or shortly *large random matrices*.

Smart Grid using Big Data Analytics: A Random Matrix Theory Approach, First Edition.
Robert C. Qiu and Paul Antonik.

If a $p \times 1$ random vector $\mathbf{X} = \left(X_1, \ldots, X_p\right)^T$ has mean vector $\boldsymbol{\mu} = \left(\mu_1, \ldots, \mu_p\right)^T$, the covariance matrix of $\mathbf{X}$ is defined by

$$\boldsymbol{\Sigma} \equiv \text{Var}\,(\mathbf{X}) = \mathbb{E}\left[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\right]$$

Furthermore, if $q \times 1$ random vector $\mathbf{Y} = \left(Y_1, \ldots, Y_q\right)^T$ with a mean vector $\boldsymbol{\eta} = \left(\eta_1, \ldots, \eta_q\right)^T$, the covariance matrix of $\mathbf{X}$ and $\mathbf{Y}$ is defined by

$$\text{Cov}\,(\mathbf{X}, \mathbf{Y}) = \mathbb{E}\left[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\eta})^T\right]$$

In particular, $\text{Cov}\,(\mathbf{X}, \mathbf{X}) = \text{Var}\,(\mathbf{X})$.

The eigenvalue decomposition is

$$\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$$

where $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}$, $\mathbf{U}$ is the unitary matrix, and $\boldsymbol{\Lambda} = \text{diag}\left(\lambda_1, \ldots, \lambda_p\right)$ is the diagonal matrix with positive eigenvalues $\lambda_1, \ldots, \lambda_p$. The generalized variance is defined as

$$\det(\boldsymbol{\Sigma}) = \det(\boldsymbol{\Lambda}) = \lambda_1 \cdots \lambda_p$$

The other overall measure is

$$\text{Tr}\,(\boldsymbol{\Sigma}) = \text{Tr}\left(\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T\right) = \text{Tr}\left(\boldsymbol{\Lambda}\mathbf{U}^T\mathbf{U}\right) = \text{Tr}\,(\boldsymbol{\Lambda}) = \lambda_1 + \cdots + \lambda_p$$

which is called the total variance. The eigenvalues of the matrix logarithm function $\log(\boldsymbol{\Sigma})$ are $\log \lambda_i$. Thus we have

$$\log\left[\det(\boldsymbol{\Sigma})\right] = \sum_{i=1}^{p} \log \lambda_i, \quad \text{and} \quad \text{Tr}\left(\log(\boldsymbol{\Sigma})\right) = \sum_{i=1}^{p} \log \lambda_i$$

So

$$\text{Tr}\left(\log(\boldsymbol{\Sigma})\right) = \log\left[\det(\boldsymbol{\Sigma})\right]$$

which is valid for any positive definite matrix $\boldsymbol{\Sigma} > 0$.

Let $\mathbf{X}$ be a p-dimensional random vector. Suppose that the probability of the random point falling in any (measurable) set E in the p-dimensional Euclidean space $\mathbb{R}^p$ is expressed as

$$\mathbb{P}(\mathbf{X} \in E) = \int_E f(\mathbf{x})\, d\mathbf{x}$$

where $d\mathbf{x} = dx_1 \cdots dx_p$. Then the function $f(\mathbf{x})$ is called the probability density function, or simply, the density of $\mathbf{X}$. The characteristic function of $\mathbf{X}$ is defined as

$$\Phi(\mathbf{t}) = \mathbb{E}\left[e^{j\mathbf{t}^T\mathbf{X}}\right]$$

where $j = \sqrt{-1}$, $\mathbf{t} = \left[t_1, \ldots, t_p\right]^T$, and $-\infty < t_i < \infty$, $i = 1, \ldots, p$. There exists one-to-one correspondence between the distribution of $\mathbf{X}$ and its characteristic function.

If the $p \times 1$ random vector $\mathbf{X}$ has the density function $f(\mathbf{x})$ and the characteristic function $\Phi(\mathbf{t})$, then

$$f(\mathbf{x}) = \frac{1}{(2\pi)^p} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-j\mathbf{t}^T\mathbf{x}} \Phi(\mathbf{t})\, dt_1 \cdots dt_p$$

B.2 Multivariate Normal Distribution

The density function of a random variable Z with the standard normal (or Gaussian) distribution $\mathcal{N}(0,1)$ is

$$f(z) \equiv \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < x < \infty.$$

A random variable X with a general normal distribution with mean μ and variance σ^2 is obtained by the linear transformation

$$X = \sigma Z + \mu$$

Thus we have

$$f(x) \equiv \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty$$

The approach is generalized as follows. The probability density function of $\mathbf{Z} = \left(Z_1, \ldots, Z_p\right)^T$, where $Z_1, \ldots, Z_p$ are independent and identically distributed (i.i.d) $\mathcal{N}(0,1)$, is given by

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} = \left(\frac{1}{\sqrt{2\pi}}\right)^p e^{-\mathbf{z}^T\mathbf{z}/2}$$

Consider the transformation

$$\mathbf{X} = \boldsymbol{\Sigma}^{1/2}\mathbf{Z} + \boldsymbol{\mu}$$

The density function is given by

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}} \frac{1}{\sqrt{\det \boldsymbol{\Sigma}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{B.1}$$

where $\boldsymbol{\mu} = \left(\mu_1, \ldots, \mu_p\right)^T$, and $\boldsymbol{\Sigma} > 0$.

Let $\boldsymbol{\mu}$ be a p-dimensional fixed vector and $\boldsymbol{\Sigma}$ be a $p \times p$ positive definite matrix. The following two statements are equivalent:

- $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$;
- $\mathbf{Z} \equiv \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \mu) \sim \mathcal{N}_p\left(\mathbf{0}, \mathbf{I}_p\right)$.

Let $\mathbf{X}_1, \ldots, \mathbf{X}_p$ be independent p-dimensional normal vectors with means $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_p$ and the same covariance matrix $\boldsymbol{\Sigma}$. Put $\mathbf{X} = \left(\mathbf{X}_1, \ldots, \mathbf{X}_n\right)^T$, and consider the transform

$$\mathbf{Y} = \left(\mathbf{Y}_1, \ldots, \mathbf{Y}_n\right)^T = \mathbf{H}\mathbf{X}$$

where $\mathbf{H}$ is an $n \times n$ orthogonal matrix. Then $\mathbf{Y}$ has the same properties as $\mathbf{X}$ except that the means of $\mathbf{Y}$ is changed to $\mathbb{E}(\mathbf{Y}) = \mathbf{H}\mathbb{E}(\mathbf{X})$.

The density function of the $n \times p$ random matrix $\mathbf{X}$ is given by

$$\begin{aligned}
&\prod_{i=1}^{n} \frac{1}{(2\pi)^{p/2}} \frac{1}{\sqrt{\det \boldsymbol{\Sigma}}} \operatorname{etr}\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu)\right\} \\
&= \frac{1}{(2\pi)^{pn/2}} \frac{1}{(\det \boldsymbol{\Sigma})^{n/2}} \operatorname{etr}\left\{-\frac{1}{2}\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\mathbf{M})^T(\mathbf{X}-\mathbf{M})\right\}
\end{aligned}$$

where etr $(\bullet)$ represents $\exp(\operatorname{Tr}(\bullet))$ and $\mathbf{X} = \left(\mathbf{X}_1, \ldots, \mathbf{X}_n\right)^T$, and $\mathbf{M} = \mathbb{E}(\mathbf{X})$.

The distribution of an $n \times p$ matrix $\mathbf{X} = \left(\mathbf{X}_{(1)}, \cdots, \mathbf{X}_{(p)}\right)$ is said to be normal if the random vector defined by

$$\text{vec}\,(\mathbf{X}) \equiv \begin{pmatrix} \mathbf{X}_{(1)} \\ \vdots \\ \mathbf{X}_{(p)} \end{pmatrix}$$

follows an np-variate normal distribution. If this is the case we simply write $\Xi = \mathbb{E}\,(\mathbf{X})$ and $\Psi = \text{Var}\,(\text{vec}\,(\mathbf{X}))$.

By using the *matrix normal distribution*, we have

$$\mathbf{X} \sim \mathcal{N}_{n\times p}\left(\Xi, \mathbf{\Sigma} \otimes \mathbf{I}_n\right) \Rightarrow \mathbf{Y} = \mathbf{H}\mathbf{X} \sim \mathcal{N}_{n\times p}\left(\mathbf{H}\Xi, \mathbf{\Sigma} \otimes \mathbf{I}_n\right)$$

where $\mathbf{H}$ is an orthogonal matrix. Here $\otimes$ denotes the Kronecker or direct product; that is, for matrices $\mathbf{A} = \left(a_{ij}\right)$ and $\mathbf{B}$, $\mathbf{A} \otimes \mathbf{B} = \left(a_{ij}\mathbf{B}\right)$.

When we write that an $r \times s$ random matrix $\mathbf{Y}$ is normally distributed, say, $\mathbf{Y}$ is $\mathcal{N}_{r\times s}\,(\mathbf{M}, \mathbf{C} \otimes \mathbf{D})$, where $\mathbf{M}$ is $r \times s$, and $\mathbf{C}$ and $\mathbf{D}$ are the positive definite matrices, we simply mean that $\mathbb{E}\,(\mathbf{Y}) = \mathbf{M}$ and that $\mathbf{C} \otimes \mathbf{D}$ is the covariance matrix of the vector $\mathbf{y} = \text{vec}\,(\mathbf{Y})$.

If the $r \times s$ matrix $\mathbf{Y}$ is $\mathcal{N}_{r\times s}\,(\mathbf{M}, \mathbf{C} \otimes \mathbf{D})$, where $\mathbf{C}\,(r \times r)$ and $\mathbf{D}\,(s \times s)$ are positive definite, then the density function of $\mathbf{Y}$ is

$$(2\pi)^{-rs/2}(\det \mathbf{C})^{-s/2}(\det \mathbf{D})^{-r/2}\,\text{etr}\left[-\frac{1}{2}\mathbf{C}^{-1}\,(\mathbf{Y}-\mathbf{M})\,\mathbf{D}^{-1}(\mathbf{Y}-\mathbf{M})^T\right]$$

B.3 Wishart Distribution

B.3.1 Central Wishart Distribution

The Wishart distribution is a multivariate generalization of the chi-square distribution. The distributions of sample covariance matrix and various sums of squares and products are Wishart, provided the underlying distribution is normal.

Let $\mathbf{X}$ be a p-dimensional random vector that is distributed as $\mathcal{N}_p\,(\boldsymbol{\mu}, \mathbf{\Sigma})$ and $\mathbf{X}_1, \cdots, \mathbf{X}_n$ are a random sample of $\mathbf{X}$ with size n. Then the sample mean vector and covariance matrices are defined by

$$\overline{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i \quad \text{and} \quad \mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)^T$$

respectively. We know $\overline{\mathbf{X}} \sim \mathcal{N}_p\,(\boldsymbol{\mu}, (1/n)\,\mathbf{\Sigma})$.

If a $p \times p$ random matrix $\mathbf{W}$ is expressed as

$$\mathbf{W} = \sum_{i=1}^{n}\mathbf{Z}_i\mathbf{Z}_i^T$$

where $\mathbf{Z}_i \sim \mathcal{N}_p\left(\boldsymbol{\mu}_i, \mathbf{\Sigma}\right)$ and $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ are independent, $\mathbf{W}$ is said to be a *noncentral Wishart* distribution with n degrees of freedom, covariance matrix $\mathbf{\Sigma}$, and noncentrality matrix $\mathbf{\Delta} = \boldsymbol{\mu}_1\boldsymbol{\mu}_1^T + \cdots + \boldsymbol{\mu}_n\boldsymbol{\mu}_n^T$. We write $\mathbf{W} \sim \mathcal{W}_p\,(n, \mathbf{\Sigma}; \mathbf{\Delta})$. In the special case of $\mathbf{\Delta} = \mathbf{0}$, we write $\mathbf{W} \sim \mathcal{W}_p\,(n, \mathbf{\Sigma})$.

If $\mathbf{A}$ is $\mathcal{W}_m(n, \mathbf{\Sigma})$ with $n \geq m$, then the density function of $\mathbf{A}$ is

$$\frac{1}{2^{mn/2}\Gamma\left(\frac{1}{2}n\right)(\det \mathbf{\Sigma})^{n/2}}(\det \mathbf{A})^{(n-m-1)/2} \operatorname{etr}\left(-\frac{1}{2}\mathbf{\Sigma}^{-1}\mathbf{A}\right) \quad (\mathbf{A} > 0)$$

where $\Gamma_m(\cdot)$ denotes the multivariate gamma function. The multivariate gamma function is defined as

$$\Gamma_m(a) = \int_{\mathbf{A}>0} \operatorname{etr}(-\mathbf{A})(\det \mathbf{A})^{a-(m+1)/2}(d\mathbf{A})$$

where $\operatorname{Re} a > \frac{1}{2}(m-1)$, and the integral is over the space of positive definite (and hence symmetric) $m \times m$ matrices. When $m = 1$, we drop m in $\Gamma_m(a)$.

If $\left(\mathbf{X}_1, \ldots, \mathbf{X}_n\right)$ are independent $\mathcal{N}_p\left(\boldsymbol{\mu}_i, \mathbf{\Sigma}\right)$ random vectors and $n > p$, the density of the sample covariance matrix

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)^T$$

is

$$\frac{1}{\Gamma\left(\frac{1}{2}n\right)(\det \mathbf{\Sigma})^{n/2}}\left(\tfrac{1}{2}n\right)^{mn/2}(\det \mathbf{S})^{(n-m-1)/2} \operatorname{etr}\left(-\frac{1}{2}n\mathbf{\Sigma}^{-1}\mathbf{S}\right) \quad (\mathbf{S} > 0)$$

The sum of independent Wishart matrices with the same covariance matrix is also Wishart. If the $m \times m$ random matrices $\mathbf{A}_1, \ldots, \mathbf{A}_N$ are all independent and $\mathbf{A}_i$ is $\mathcal{W}_m(n, \mathbf{\Sigma})$, $i = 1, \ldots, N$, then $\sum_{i=1}^{N}\mathbf{A}_i$ is also $\mathcal{W}_m(n, \mathbf{\Sigma})$, where $n = \sum_{i=1}^{N} n_i$.

B.3.2 Noncentral Wishart Distribution

The noncentral Wishart distribution generalizes the noncentral χ^2 distribution in the same way that the usual or central Wishart distribution generalizes the χ^2 distribution. It forms a major block for noncentral distributions.

If $\mathbf{A} = \mathbf{Z}^T\mathbf{Z}$, where the $n \times m$ matrix $\mathbf{Z}$ is $\mathcal{N}_{n\times m}\left(\mathbf{M}, \mathbf{I}_n \otimes \mathbf{\Sigma}\right)$, then $\mathbf{A}$ is said to have the noncentral Wishart distribution with n degrees of freedom, covariance matrix $\mathbf{\Sigma}$, and matrix of noncentrality parameters $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}\mathbf{M}^T\mathbf{M}$. We will write that $\mathbf{A}$ is $\mathcal{W}_m(n, \mathbf{\Sigma}; \mathbf{\Omega})$.

If the $n \times m$ matrix $\mathbf{Z}$ is $\mathcal{N}_{n\times m}\left(\mathbf{M}, \mathbf{I}_n \otimes \mathbf{\Sigma}\right)$, with $n \geq m$, then the density function of $\mathbf{A} = \mathbf{Z}^T\mathbf{Z}$ is

$$\frac{1}{2^{mn/2}\Gamma_m\left(\frac{1}{2}n\right)(\det \mathbf{\Sigma})^{n/2}}(\det \mathbf{A})^{(n-m-1)/2} \operatorname{etr}\left(-\frac{1}{2}\mathbf{\Omega}\right) {}_0F_1$$
$$\left(\tfrac{1}{2}n; \tfrac{1}{4}\mathbf{\Omega}\mathbf{\Sigma}^{-1}\mathbf{A}\right) \quad (\mathbf{A} > 0)$$

where $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}\mathbf{M}^T\mathbf{M}$. Here the hypergeometric function of matrix argument ${}_0F_1(a; \mathbf{X})$ is defined as

$${}_0F_1(a; \mathbf{X}) = \det(\mathbf{I} - \mathbf{X})^a$$

The sum of independent noncentral Wishart matrices with the same covariance matrix is also noncentral Wishart. If the $m \times m$ matrices $\mathbf{A}_1, \ldots, \mathbf{A}_N$ are all independent

and $\mathbf{A}_i$ is $\mathcal{W}_m\left(n_i, \mathbf{\Sigma}; \mathbf{\Omega}_i\right), i = 1, \ldots, N$, then $\sum_{i=1}^{N} \mathbf{A}_i$ is also $\mathcal{W}_m(n, \mathbf{\Sigma}; \mathbf{\Omega})$, with $n = \sum_{i=1}^{N} n_i$ and $\mathbf{\Omega} = \sum_{i=1}^{N} \mathbf{\Omega}_i$.

If the $n \times m$ matrix Z is $\mathcal{N}_{n\times m}\left(\mathbf{M}, \mathbf{I}_n \otimes \mathbf{\Sigma}\right)$, and $\mathbf{Q}$ is $k \times m$ of rank k, then $\mathbf{QZ}^T\mathbf{ZQ}^T$ is $\mathcal{W}_k\left(n, \mathbf{Q\Sigma Q}^T; \left(\mathbf{Q\Sigma Q}^T\right)^{-1}\mathbf{QM}^T\mathbf{MQ}^T\right)$.

If $\mathbf{A}$ is $\mathcal{W}_m(n, \mathbf{\Sigma}; \mathbf{\Omega})$, with $n \geq m$

$$\mathbb{E}\left[(\det \mathbf{A})^r\right] = (\det \mathbf{\Sigma})\, 2^{mr} \frac{\Gamma_m\left(\frac{1}{2}n + r\right)}{\Gamma_m\left(\frac{1}{2}n\right)} {}_1F_1\left(-r; \tfrac{1}{2}n; -\tfrac{1}{2}\mathbf{\Omega}\right)$$

Note that this is a polynomial of degree mr if r is a positive integer. Here the "confluent" hypergeometric function ${}_1F_1$ is defined as

$${}_1F_1(a; c; \mathbf{X}) = \frac{\Gamma_m(c)}{\Gamma_m(a)\,\Gamma_m(c-a)} \int_{0<\mathbf{Y}<\mathbf{I}_m} \operatorname{etr}(\mathbf{XY})$$
$$(\det \mathbf{Y})^{a-(m+1)/2} \det(\mathbf{I} - \mathbf{Y})^{c-a-(m+1)/2} (d\mathbf{Y})$$

valid for all symmetric $\mathbf{X}$, $\operatorname{Re}(a) > \frac{1}{2}(m-1), \operatorname{Re}(c) > \frac{1}{2}(m-1)$, and $\operatorname{Re}(c-a) > \frac{1}{2}(m-1)$.

B.4 Multivariate Linear Model

The multivariate linear model is ubiquitous; for example, it is such as in the MIMO and state estimation. We are interested in high-dimensional applications. The multivariate linear model allows a vector of observations, given by the rows of a matrix $\mathbf{Y}$, to correspond to the rows of the known matrix $\mathbf{H}$. The multivariate linear model takes the form

$$\mathbf{Y} = \mathbf{HX} + \mathbf{N} \tag{B.2}$$

where $\mathbf{Y}$ and $\mathbf{N}$ are $m \times m$ random matrices, $\mathbf{H}$ is a known $n \times p$ matrix, and $\mathbf{X}$ is an unknown $p \times m$ matrix of parameters called regression coefficients. We assume throughout this section that $\mathbf{H}$ has rank p, that $n \geq m + p$, and the rows of the noise matrix $\mathbf{N}$ are independent $\mathcal{N}_m(\mathbf{0}, \mathbf{\Sigma})$ random vectors. Using the notation introduced above, this means that $\mathbf{N}$ is $\mathcal{N}_{n\times m}\left(\mathbf{0}, \mathbf{I}_n \otimes \mathbf{\Sigma}\right)$, so that $\mathbf{Y}$ is $\mathcal{N}_{n\times m}\left(\mathbf{HX}, \mathbf{I}_n \otimes \mathbf{\Sigma}\right)$.

We now find the maximum likelihood estimates of $\mathbf{X}$ and $\mathbf{\Sigma}$ and show that they are sufficient.

Theorem B.4.1 If $\mathbf{Y}$ is $\mathcal{N}_{n\times m}\left(\mathbf{HX}, \mathbf{I}_n \otimes \mathbf{\Sigma}\right)$, and $n \geq m + p$ the maximum likelihood estimates of $\mathbf{X}$ and $\mathbf{\Sigma}$ are

$$\hat{\mathbf{X}} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{Y} \tag{B.3}$$

and

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\left(\mathbf{Y} - \mathbf{H}\hat{\mathbf{X}}\right)^T\left(\mathbf{Y} - \mathbf{H}\hat{\mathbf{X}}\right) \tag{B.4}$$

Moreover, $\left(\hat{\mathbf{X}}, \hat{\mathbf{\Sigma}}\right)$ is sufficient for $\left(\mathbf{X}, \mathbf{\Sigma}\right)$.

If $\mathbf{Y}$ is $\mathcal{N}_{n\times m}\left(\mathbf{HX}, \mathbf{I}_n \otimes \mathbf{\Sigma}\right)$, the maximum likelihood estimates of $\mathbf{X}$ and $\mathbf{\Sigma}$ are given by (B.3) and (B.4), respectively. Then, $\hat{\mathbf{X}}$ and $\hat{\mathbf{\Sigma}}$ are independently distributed; $\hat{\mathbf{X}}$ is $\mathcal{N}_{p\times m}\left(\mathbf{X}, \left(\mathbf{H}^T\mathbf{H}\right)^{-1} \otimes \mathbf{\Sigma}\right)$ and $n\hat{\mathbf{\Sigma}}$ is $\mathcal{W}_m\left(n-p, \mathbf{\Sigma}\right)$.

B.5 General Linear Hypothesis Testing

In this section, we consider testing the hypothesis testing

$$\mathcal{H}_0 : \mathbf{CX} = 0$$
$$\mathcal{H}_1 : \mathbf{CX} \neq 0$$

where $\mathbf{C}$ is a known $r \times p$ matrix of rank r. We partition $\mathbf{X}$ as

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$$

where $\mathbf{X}_1$ is $r \times m$ and $\mathbf{X}_2$ is $(p-r) \times m$. The null hypothesis $\mathbf{X}_1 = \mathbf{0}$ is the same as $\mathbf{CX} = \mathbf{0}, \mathbf{C} = \left[\mathbf{I}_r; \mathbf{0}\right]$.

By transforming the variables and parameters in the model $\mathbf{Y} = \mathbf{HX} + \mathbf{N}$, the problem in the transformed domain can be assumed to be following form:

$$\tilde{\mathbf{Y}} = \begin{bmatrix} \tilde{\mathbf{Y}}_1 \\ \tilde{\mathbf{Y}}_2 \\ \tilde{\mathbf{Y}}_3 \end{bmatrix}$$

where $\tilde{\mathbf{Y}}_1$ is $r \times m$, $\tilde{\mathbf{Y}}_2$ is $(p-r) \times m$, and $\tilde{\mathbf{Y}}_3$ is $(n-p) \times m$. Let $\tilde{\mathbf{Y}}$ be a random matrix whose rows are independent m-variate normal with common covariance matrix $\mathbf{\Sigma}$ and expectations give by

$$\mathbb{E}\left(\tilde{\mathbf{Y}}_1\right) = \mathbf{M}_1, \quad \mathbb{E}\left(\tilde{\mathbf{Y}}_2\right) = \mathbf{M}_2, \quad \mathbb{E}\left(\tilde{\mathbf{Y}}_1\right) = \mathbf{0}$$

The null hypothesis $\mathcal{H}_0 : \mathbf{CX} = 0$, is equivalent to $\mathcal{H}_0$: $\mathbf{M}_1 = \mathbf{0}$.

Theorem B.5.1 The likelihood ratio test of size α of $\mathcal{H}_0$: $\mathbf{M}_1 = \mathbf{0}$ against $\mathcal{H}_1$: $\mathbf{M}_1 \neq \mathbf{0}$ rejects $\mathcal{H}_0$ if $\Lambda \leqslant c_\alpha$, where

$$W = \Lambda^{2/n} = \frac{\det \mathbf{B}}{\det\left(\mathbf{A} + \mathbf{B}\right)}$$

with $\mathbf{A} = \tilde{\mathbf{Y}}_1^T\tilde{\mathbf{Y}}_1$ is $\mathcal{W}_m\left(r, \mathbf{\Sigma}; \mathbf{\Omega}\right)$, $\mathbf{B} = \tilde{\mathbf{Y}}_3^T\tilde{\mathbf{Y}}_3$ is $\mathcal{W}_m\left(n-p, \mathbf{\Sigma}\right)$, $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}\mathbf{M}_1^T\mathbf{M}_1$, and c_α is chosen so that the size of the test is α.

The likelihood ratio test is equivalent to rejecting $\mathcal{H}_0 : \mathbf{M}_1 = \mathbf{0}$ for small values of $W = \Lambda^{2/n}$. This is an invariant test for

$$\begin{aligned} W &= \frac{\det \mathbf{B}}{\det\left(\tilde{\mathbf{Y}}_1^T\tilde{\mathbf{Y}}_1 + \mathbf{B}\right)} \\ &= \det\left(\mathbf{I} + \tilde{\mathbf{Y}}_1^T\mathbf{B}^{-1}\tilde{\mathbf{Y}}\right)^{-1} \\ &= \prod_{i=1}^{s}\left(1+\lambda_i\right)^{-1} \end{aligned}$$

where $s = \min(r, m) = \text{rank}\left(\tilde{\mathbf{Y}}_1^T \mathbf{B}^{-1} \tilde{\mathbf{Y}}\right)$ and $\lambda_1 \geqslant \cdots \geqslant \lambda_s > 0$ are the nonzero eigenvalues of $\tilde{\mathbf{Y}}_1^T \mathbf{B}^{-1} \tilde{\mathbf{Y}}$. Or we consider

$$-\log W = \sum_{i=1}^{s} \log\left(1 + \lambda_i\right)$$

which is in the form of linear eigenvalue statistics. Central limit theorem for linear eigenvalue statistics of sample covariance random matrices [214, 215] can be used. We refer to Section 3.7 for details. See also [474] and references there.

When the dimensions of the matrices are high, concentration of spectral measure phenomenon occurs. We refer to Qiu and Wicks [40] for details. The statistic W^h can be shown to be highly concentrated around some value (most often its expectation). General MANOVA matrices [456] are an extension of the multivariate analysis of variance to determine correlation coefficients (Section 3.3 of [37]).

The h-th moment of W, when $n - p \geq m$, $r \geq m$, is

$$\mathbb{E}\left(W^h\right) = \frac{\Gamma_m(\frac{1}{2}(n-p)+h)\Gamma_m\left(\frac{1}{2}(n+r-p)\right)}{\Gamma_m\left(\frac{1}{2}(n-p)\right)\Gamma_m\left(\frac{1}{2}(n+r-p)+h\right)} F_1\left(h; \tfrac{1}{2}(n+r-p)+h; \tfrac{1}{2}\mathbf{\Omega}\right)$$

where $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}\mathbf{M}_1^T\mathbf{M}_1$, for $r \geq m$.

The moments of W when $\mathcal{H}_0 : \mathbf{M}_1 = \mathbf{0}$ are obtained by putting $\mathbf{\Omega} = \mathbf{0}$. In this case, the moments of W are given by

$$\mathbb{E}\left(W^h\right) = \frac{\Gamma_m\left(\frac{1}{2}(n-p)+h\right)\Gamma_m\left(\frac{1}{2}(n+r-p)\right)}{\Gamma_m\left(\frac{1}{2}(n-p)\right)\Gamma_m\left(\frac{1}{2}(n+r-p)+h\right)}$$

Bibliographical Remarks

Classical multivariate analysis provides a good starting point for us to become familiar with matrix-valued random variables. We extract some of the most relevant facts from classical texts. More details can be found from the classical texts: Scrivastava and Khatri [659], Siotani *et al.* (1985) [660], Anderon [371], Muirhead [37], and Fujikoshi *et al.* [483]. The role of this appendix has been to provide some preliminary materials. We make no attempt to make this appendix comprehensive. At most, we want to give readers a feel of the key mathematical concepts in the classical settings.

The general motivation is to study large random matrices. We have tried our best to make the book self-contained. One general application is for hypothesis testing in high dimensions. See [474] and references there for some applications. We need to revisit some classical algorithms. For the dimension p of the dataset and the sample size n, the classical algorithms make the assumption that p is a fixed small constant or at least negligible compared with the sample size n. This assumption is no longer true for many modern datasets, such as big data, because their dimensions can be proportionally large compared with the sample size. For example, financial data, consumer data, sensors, data, the communication network data, the smart grid data, the modern manufacturing data, and the multimedia data all have this feature.

Most contents of this book, together with its two companion books [39, 40], are beyond the scope of the above classical books. Our three books are founded upon two themes: (i) random matrix theory; (ii) concentration of spectral measures. The first theme deals with statistics when the dimensions of random matrices are asymptotically large—both n and p go to infinity at the same rate. The second theme, on the other hand, deals with an nonasymptotic analysis of random matrices—both n and p are large but finite!

Our three books are, in some sense, complementary to the listed classical books. Our books are structured using mathematics. From our point of view, big data is a statistical science that uses large random matrices to model the datasets.

Appendix B is adapted from [292, 661].

References

1 Z. Burda, J. Kornelsen, M. A. Nowak *et al.* (2013) "Collective correlations of brodmann areas fmri study with rmt-denoising," *arXiv preprint arXiv:1306.3825*.

2 A. Halevy, P. Norvig, and F. Pereira (2009) The unreasonable effectiveness of data. *Intelligent Systems, IEEE*, **24**(2), 8–12.

3 E. P. Wigner (1960) The unreasonable effectiveness of mathematics in the natural sciences. Richard Courant Lecture in Mathematical Sciences delivered at New York University, May 11, 1959. *Communications on Pure and Applied Mathematics* **13**(1), 1–14.

4 Z. Tian (2013) Big Data: From Signal Processing to Systems Engineering in NSF Workshop on Big Data: From Signal Processing to Systems Engineering, Arlington, VA, March.

5 NSF (2012) Core Techniques and Technologies for Advancing Big Data Science and Engineering (Bigdata), *NSF technical report.*

6 World Economic Forum (2012) *Big Data Big Impact: New Possibilities for International Development*, http://www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_Briefing_2012.pdf (accessed September 20, 2016).

7 Big data (2008) *Nature*, http://www.nature.com/news/specials/bigdata/index.html (accessed September 11, 2016).

8 Data, data everywhere (2010) The *Economist*, February 25.

9 Drowning in numbers—digital data will flood the planet–and help us understand it better (2011) *The Economist*, November 18.

10 D. Agrawal, P. Bernstein, E. Bertino, *et al.*, Challenges and opportunities with big data (2012) *Proceedings of the VLDB Endowment*, vol. 5, no. 12, pp. 2032–2033 http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf (accessed September 11, 2016).

11 S. Lohr (2012) The age of big data." *New York Times*, February 12.

12 J. Manyika, M. Chui, B. Brown, *et al*. (2011), Big data: the next frontier for innovation. *McKinsey Global Institute report*, May.

13 Y. Noguchi, (2011) Following digital breadcrumbs to big data gold. National Public Radio, November 29.

14 Big data (2013) *New York Times* (special section on the business and culture of big data), June 25.

15 J. Chen, Y. Chen, X. Du, *et al.* (2013) Big data challenge: a data management perspective *Frontiers of Computer Science* **7**(2), vol. 7, no. 2, pp. 157–164.

16 Big data, (2011), *Science* (special section), http://www.sciencemag.org/site/special/data/ (accessed September 11, 2016).

Smart Grid using Big Data Analytics: A Random Matrix Theory Approach, First Edition.
Robert C. Qiu and Paul Antonik.

17 T. Kalil (2012) *Big Data is a Big Deal*, https://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal (accessed September 20, 2016).
18 T. H. Davenport, P. Barth, and R. Bean (2012) How big data is different, *MIT Sloan Management Review* **54**(1), 22–24.
19 DARPA (2013) *Extracting Relevance from Mountains of Data*, https://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal (accessed September 20, 2016).
20 D. Pedreschi (2013) *The Future of Big Data*, www.storify.com/katarzynasz/the-future-of-big-data (accessed September 20, 2016).
21 C. Anderson (2008) Will the data deluge make the scientific method obsolete? *Edge* June.
22 G. Li (2012) The scientific value of big data, *Research Communications of the Chinese Computer Society* **8**(9), 8–15.
23 DARPA (2012) Broad agency announcement xdata, *DARPA tech. rep. DARPA-BAA-12-13*.
24 A. Labrinidis and H. Jagadish (2012) Challenges and opportunities with big data, *Proceedings of the VLDB Endowment*, **5**(12), 2032–2033.
25 J. Dean and S. Ghemawat (2008) Mapreduce: simplified data processing on large clusters. *Communications of the ACM* **51**(1), pp. 107–113.
26 A. Hero (2013) Winnowing signals from massive data: Sp for big data and its relation to systems engineering, in NSF Workshop on Big Data: From Signal Processing to Systems Engineering, Arlington, VA, March.
27 R. G. Baraniuk (2011) More is less: signal processing and the data deluge. *Science (Washington)* **331**(6018), pp. 717–719.
28 S. Ganguli and H. Sompolinsky (2012) Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis. *Annual Review of Neuroscience* **35**, 485–508.
29 J. Calder, S. Esedoglu, and A. O. Hero (2013) A Hamilton–Jacobi equation for the continuum limit of non-dominated sorting, *arXiv preprint arXiv:1302.5828*.
30 K.-J. Hsiao, K. S. Xu, J. Calder, and A. O. Hero III (2011) Multi-criteria anomaly detection using Pareto depth analysis, *arXiv preprint arXiv:1110.3741*.
31 R. C. Qiu (2012) *Towards a Large-Scale Cognitive Radio Network: Testbed, Distributed Sensing, and Random Matrices*, technical report. Research Proposal to NSF, Tennessee Technological University, February.
32 R. Qiu (2012) *Collection, Analysis and Exploitation of Big Data in Cognitive Radio Networks*, technical report. Research Proposal to NSF, Tennessee Technological University, November.
33 R. Qiu (2012) *Collaborative Research: Towards a Large-Scale Heterogeneous Cognitve Radio Network System: Tests and Validation, Big Data, and Cognitive Spectrum Management*, technical report. Research Proposal to NSF, Tennessee Technological University, June.
34 R. C. Qiu, (2012) *Towards a Large-Scale Cognitive Radio Network: Testbed, Distributed Sensing, and Random Matrices*, technical report. Proposal to NSF, Tennessee Technological University.
35 G. W. Anderson, A. Guionnet, and O. Zeitouni (2010) *An Introduction to Random Matrices*, Cambridge University Press.
36 K. Mardia, J. Kent, and J. Bibby (1979) *Multivariate Analysis*, Academic Press.
37 R. Muirhead (2005) *Aspects of Mutivariate Statistical Theory*, John Wiley & Sons, Ltd.

38 D. Paul and A. Aue (2013) Random matrix theory in statistics: a review. *Journal of Statistical Planning and Inference* **150**, 1–29.

39 R. Qiu, Z. Hu, H. Li, and M. Wicks (2012) *Cognitive Communications and Networking: Theory and Practice*, John Wiley & Sons, Ltd.

40 R. Qiu and M. Wicks (2013) *Cognitive Networked Sensing and Big Data*, Springer Verlag.

41 T. Kolda (2013) *Matlab Tensor Toolbox*, version 2.5.

42 *Big Data across the Federal Government* (2012) Technical report. Executive Office of the President, March.

43 A. Rajaraman and J. D. Ullman (2012) *Mining of Massive Datasets,* Cambridge University Press.

44 C. Zhang and R. C. Qiu (2014) Data modeling with large random matrices in a cognitive radio network testbed: Initial experimental demonstrations with 70 nodes, *arXiv preprint arXiv:1404.3788*.

45 X. Wu, X. Zhu, G.-Q. Wu, and W. Ding (2014) Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering* **26**(1), 97–107.

46 S. Barbarossa, S. Sardellitti, and P. Di Lorenzo (2013) Distributed detection and estimation in wireless sensor networks, *arXiv preprint arXiv:1307.1448*.

47 P. E. Dewdney, P. J. Hall, R. T. Schilizzi, and T. J. L. Lazio (2009) The square kilometre array. *Proceedings of the IEEE* **97**(7), 1482–1496.

48 E. Birney (2012) The making of encode: lessons for big-data projects. *Nature* **489**(7414), 49–51.

49 S. Boucheron, G. Lugosi, and P. Massart (2013) *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press.

50 M. Talagrand (1995) Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathematiques de l'IHES*, **81**(1), 73–205.

51 E. Telatar (1999) Capacity of multi-antenna gaussian channels. *European Transactions on Telecommunications*, **10**(6), 585–595.

52 A. Tulino and S. Verdu, (2004) *Random Matrix Theory and Wireless Communications*, Now Publishers Inc.

53 Z. Bai, J. Chen, and J. Yao (2010) On estimation of the population spectral distribution from a high-dimensional sample covariance matrix. *Australian and New Zealand Journal of Statistics*, **52**(4), 423–437.

54 Tucci, G. H. and Vega, M. V. (2013) A note on functional averages over Gaussian ensembles. *Journal of Probability and Statistics*, https://www.hindawi.com/journals/jps/2013/941058/ (accessed September 20, 2016).

55 T. Tao, V. Vu, and M. Krishnapur (2010) Random matrices: Universality of ESDS and the circular law. *The Annals of Probability*, **38**(5), 2023–2065.

56 D. Chafaï (2014) From Boltzmann to random matrices and beyond, *arXiv preprint arXiv:1405.1003*.

57 Chafai, D., Gozlan, N. and Zitt, P. A. (2013) First order global asymptotics for confined particles with singular pair repulsion. *Annals of Applied Probability* **24**(6), 2371–2431.

58 D. Petz, (2001) Entropy, von Neumann and the Von Neumann entropy, in *John von Neumann and the Foundations of Quantum Physics* (eds. M. Redei and M. Stöltzner). Kluwer, pp. 83–96.

59 T. Cover and J. Thomas (2006) *Elements of Information Theory*. John Wiley & Sons, Ltd.

60 F. Li, W. Qiao, H. Sun, *et al.* (2010) Smart transmission grid: Vision and framework. *IEEE Transactions on Smart Grid* **1**(2), 168–177.

61 F. Lin, R. C. Qiu, Z. Hu, *et al.* (2012) Generalized fmd detection for spectrum sensing under low signal-to-noise ratio. *Communications Letters, IEEE* **16**(5), 604, 607.

62 P. J. Forrester, (2010) *Log-Gases and Random Matrices (LMS-34)*, Princeton University Press.

63 J. H. Porter, P. C. Hanson, and C.-C. Lin (2012) Staying afloat in the sensor data deluge. *Trends in Ecology and Evolution* **27**(2), 121–129.

64 D. Thompson, S. Burke-Spolaor, A. Deller, *et al.* (2013) Real time adaptive event detection in astronomic data streams: Lessons from the very long baseline array. *IEEE Intelligent Systems* **29**(1), 48–55.

65 D. L. Jones (2013) Technology challenges for the square kilometer array. *IEEE Aerospace and Electronics Systems Magazine* **28**(2) 18–23.

66 A.-J. van der Veen and S. J. Wijnholds (2013) Signal processing tools for radio astronomy, in *Handbook of Signal Processing Systems* (eds S.S. Bhattacharyya and E.F. Deprettere). Springer, pp. 421–463.

67 T. Tao (2012) *Topics in Random Matrix Thoery*, American Mathematical Society.

68 R. A. Fisher (1922) On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, pp. 309–368.

69 N. L. Johnson (1993) *Breakthroughs in Statistics: Foundations and Basic Theory, Volume I*, vol. 1. Springer.

70 X. Li, F. Lin, and R. C. Qiu, (2014) Modeling massive amount of experimental data with large random matrices in a real-time uwb-mimo system, *arXiv preprint arXiv:1404.4078*.

71 R. Speicher (2014) Free probability and random matrices, *arXiv preprint arXiv:1404.3393*.

72 Z. Burda, R. Janik, and B. Waclaw (2010) Spectrum of the product of independent random gaussian matrices. *Physical Review E* **81**(4), 041132.

73 Z. Burda, A. Jarosz, G. Livan, *et al.* (2010) Eigenvalues and singular values of products of rectangular Gaussian random matrices. *Physical Review E* **82**(6), p. 061114.

74 A. Jarosz (2011) Summing free unitary random matrices. *Physical Review E*, **84**(1), 011146.

75 P. J. Forrester (2014) Eigenvalue statistics for product complex wishart matrices, *arXiv preprint arXiv:1401.2572*.

76 A. B. Kuijlaars and D. Stivigny (2014) Singular values of products of random matrices and polynomial ensembles, *arXiv preprint arXiv:1404.5802*.

77 E. Strahov (2014) Differential equations for singular values of products of ginibre random matrices, *arXiv preprint arXiv:1403.6368*.

78 P. J. Forrester and D.-Z. Liu (2014) Raney distributions and random matrix theory, *arXiv preprint arXiv:1404.5759*.

79 T. T. Cai, T. Liang, and H. H. Zhou (2013) Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional gaussian distributions, *arXiv preprint arXiv:1309.0482*.

80 Y. Ahmadian, F. Fumarola, and K. D. Miller (2013) Properties of networks with partially structured and partially random connectivity, *arXiv preprint arXiv:1311.4672*.

81 N. E. Karoui and H.-T. Wu (2013) Vector diffusion maps and random matrices with random blocks, *arXiv preprint arXiv:1310.0188*.

82 J.-P. Bouchaud, L. Laloux, M. A. Miceli, and M. Potters (2007) Large dimension forecasting models and random singular value spectra. *The European Physical Journal B* **55**(2), 201–207.

83 P. Russom (2011) Big data analytics. *TDWI Best Practices Report, Fourth Quarter*.

84 F. Bach, H. K. Çakmak, H. Maass, and U. Kuehnapfel (2013) *Power Grid Time Series Data Analysis with Pig on a Hadoop Cluster Compared to Multi Core Systems*, Twenty-First Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP), IEEE, pp. 208–212.

85 J. Depablos, V. Centeno, A. G. Phadke, and M. Ingram (2004) *Comparative Testing of Synchronized Phasor Measurement Units*, Power Engineering Society General Meeting, IEEE, pp. 948–954.

86 B. Blaszczyszyn, M. Jovanovic, and M. K. Karray (2013) Quality of real-time streaming in wireless cellular networks-stochastic modeling and analysis, *arXiv preprint arXiv:1304.5034*.

87 A. Sengupta and P. Mitra (1999) Distributions of singular values for some random matrices. *Physical Review E* **60**(3), 3389.

88 I. T. Jolliffe, N. T. Trendafilov, and M. Uddin (2003) A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics* **12**(3), 531–547.

89 A. Amini (2009) High-dimensional analysis of semidefinite relaxations for sparse principal components. *The Annals of Statistics,* **37**(5B), 2877–2921.

90 G. I. Allen, L. Grosenick, and J. Taylor (2011) A generalized least squares matrix decomposition, *arXiv preprint arXiv:1102.3074*.

91 G. McLachlan and D. Peel (2004) *Finite Mixture Models*. John Wiley & Sons, Inc.

92 A. Khalili and J. Chen (2007) Variable selection in finite mixture of regression models. *Journal of the American Statistical Association* **102**(479), 1025–1038.

93 N. Städler, P. Bühlmann, and S. van de Geer (2009) *L1 Penalization for Mixture Regression Models* ftp://ftp.stat.math.ethz.ch/Research-Reports/Other-Manuscripts/buhlmann/stadbuhlgeer-final.pdf (accessed September 12, 2016)

94 N. Städler and P. Bühlmann (2012) Missing values: sparse inverse covariance estimation and an extension to sparse regression. *Statistics and Computing* **22**(1), 219–235.

95 D. L. Donoho (2000) High-dimensional data analysis: The curses and blessings of dimensionality. AMS Math Challenges Lecture, pp. 1–32.

96 T. Hastie, R. Tibshirani, and J. Friedman (2009) *The Elements of Statistical Learning*, Springer.

97 P. L. Bühlmann, S. A. van de Geer, and S. Van de Geer (2011) *Statistics for High-Dimensional Data*. Springer.

98 J. Fan and Y. Fan (2008) High dimensional classification using features annealed independence rules. *Annals of Statistics* **36**(6), 2605.

99 J. Fan and J. Lv (2008) Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(5), 849–911.

100 S. Boyd, N. Parikh, E. Chu, *et al.* (2011) Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning* **3**(1), 1–122.

101 Fodor, I. K. (2002) A survey of dimension reduction techniques. *Lawrence Livermore National Laboratory tech. rep.*

102 E. P. Wigner (1951) On a class of analytic functions from the quantum theory of collisions. *The Annals of Mathematics* **53**(1), 36–67.

103 M. Mehta (2004) *Random Matrices*, Academic Press.

104 T. A. Brody, J. Flores, J. B. French, *et al.* (1981) Random-matrix physics: Spectrum and strength fluctuations. *Reviews of Modern Physics* **53**(3), 385.

105 T. Guhr, A. Müller-Groeling, and H. Weidenmüller (1998) Random-matrix theories in quantum physics: Common concepts. *Physics Reports*, **299**(4), 189–425.

106 M. Santhanam and P. K. Patra (2001) Statistics of atmospheric correlations. *Physical Review E* **64**(1), 016102.

107 L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters (1999) Noise dressing of financial correlation matrices. *Physical Review Letters* **83**(7), 1467.

108 N. Silver (2012) *The Signsi and the Noise: Why So Many Predictions Fail—but Some Don't*. Penguin Press.

109 E. Wigner (1958) On the distribution of the roots of certain symmetric matrices. *The Annals of Mathematics* **67**(2), 325–327.

110 E. Wigner (1955) Characteristic vectors of bordered matrices with infinite dimensions. *The Annals of Mathematics* **62**(3), 548–564,.

111 J. Ginibre (1965) Statistical ensembles of complex, quaternion, and real matrices. *Journal of Mathematical Physics* **6**, 440.

112 N. Lehmann and H.-J. Sommers (1991) Eigenvalue statistics of random real matrices. *Physical review letters* **67**(8), 941.

113 A. Edelman (1997) The probability that a random real gaussian matrix has k real eigenvalues, related distributions, and the circular law. *Journal of Multivariate Analysis* **60**(2), 203–232.

114 E. Kanzieper and G. Akemann (2005) Statistics of real eigenvalues in ginibres ensemble of random real matrices. *Physical Review Letters* **95**(23), 230201.

115 C. Biely and S. Thurner (2008) Random matrix ensembles of time-lagged correlation matrices: Derivation of eigenvalue spectra and analysis of financial time-series. *Quantitative Finance* **8**(7), 705–722.

116 J. Wishart (1928) The generalised product moment distribution in samples from a normal multivariate population. *Biometrika* **20**(1/2), 32–52.

117 V. Plerou, P. Gopikrishnan, B. Rosenow, *et al.* (1999) Universal and nonuniversal properties of cross correlations in financial time series. *Physical Review Letters* **83**(7), 1471–1474.

118 K. Mayya and R. Amritkar (2006) Analysis of delay correlation matrices, *arXiv preprint cond-mat/0601279*.

119 L. Laloux, P. Cizeau, M. Potters, and J. Bouchaud (2000). Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance* **3**(3), 391–398.

120 T. Guhr and B. Kälber (2003) A new method to estimate the noise in financial correlation matrices. *Journal of Physics A: Mathematical and General* **36**(12), 3009.

121 Z. Burda and J. Jurkiewicz (2004) Signal and noise in financial correlation matrices. *Physica A: Statistical Mechanics and its Applications* **344**(1), 67–72.

122 Z. Burda, J. Jurkiewicz, and B. Waclaw (2005) Spectral moments of correlated wishart matrices. *Physical Review E* **71**(2), 26111.

123 J. Feinberg and A. Zee (1997) Non-gaussian non-hermitian random matrix theory: phase transition and addition formalism. *Nuclear Physics B* **501**(3), 643–669.

124 Z. Burda, A. T. Gorlich, and B. Waclaw (2006) Spectral properties of empirical covariance matrices for data with power-law tails. *Physical Review E*, **74**(4), 041129.

125 Z. Burda, A. Jarosz, M. A. Nowak, *et al.* (2011) Applying free random variables to random matrix analysis of financial data. part i: The gaussian case. *Quantitative Finance* **11**(7), 1103–1124.

126 D. Voiculescu, K. Dykema, and A. Nica (1992) *Free Random Variables*. American Mathematical Society.

127 R. Gopakumar and D. J. Gross (1995) Mastering the master field. *Nuclear Physics B* **451**(1), 379–415.

128 A. Zee (1996) Law of addition in random matrix theory. *Nuclear Physics B* **474**(3), 726–744.

129 R. Janik, M. Nowak, G. Papp, and I. Zahed (1999) Localization transitions from free random variables. *Acta Physica Polonica B* **30**, 45.

130 R. Speicher (1998) Combinatorial theory of the free product with amalgamation and operator-valued free probability theory. *Memoirs of the American Mathematical Society* **133**(634), 627–627.

131 R. Janik, M. Nowak, G. Papp, *et al.* (1997) Non-hermitian random matrix models: Free random variable approach. *Physical Review E* **55**(4), 4100.

132 H. Bercovici and D. Voiculescu (1993) Free convolution of measures with unbounded support. *Indiana University Mathematics Journal* **42**(3), 733–774.

133 H. Bercovici and V. Pata (1996) The law of large numbers for free identically distributed random variables. *The Annals of Probability*, 453–465.

134 Z. Burda, J. Jurkiewicz, M. A. Nowak, *et al.* (2001) Levy matrices and financial covariances, *arXiv preprint cond-mat/0103108*.

135 J. Silverstein and Z. Bai (1995) On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate analysis*, **54**(2), 175–192.

136 R. Couillet and M. Debbah (2011) *Random Matrix Methods for Wireless Communications*. Cambridge University Press.

137 A. Guionnet, M. Krishnapur, and O. Zeitouni (2009) The single ring theorem, *arXiv preprint arXiv:0909.2214*.

138 F. Benaych-Georges and J. Rochet (2013) Outliers in the single ring theorem, *arXiv preprint arXiv:1308.3064*.

139 B. Cakmak (2012) Non-hermitian random matrix theory for mimo channels, master's thesis, Norwegian University of Science and Technology.

140 F. Boccardi, R. W. Heath Jr, A. Lozano, *et al.* (2014) Five disruptive technology directions for 5g. *IEEE Communications Magazine* **52**, 74–80.

141 C.-X. Wang, F. Haider, X. Gao, *et al.* (2014) Cellular architecture and key technologies for 5g wireless communication networks. *IEEE Communications Magazine* **52**(2), 122–130.

142 R. Bryant, R. H. Katz, and E. D. Lazowska (2008) Big-data computing: Creating revolutionary breakthroughs in commerce. 2008.

143 S. P. Ahuja and B. Moore (2013) State of big data analysis in the cloud. *Network and Communication Technologies* **2**(1), p62.

144 E. Begoli and J. Horey (2012) Design Principles for Effective Knowledge Discovery from Big Data, in *Software Architecture (WICSA) and European Conference on Software Architecture (ECSA), 2012 Joint Working IEEE/IFIP Conference on*, IEEE, 215–218.

145 G. B. Giannakis, V. Kekatos, N. Gatsis, *et al.* (2013) Monitoring and optimization for power grids: A signal processing perspective, *arXiv preprint arXiv:1302.0885*.

146 R. Davis, O. Pfaffel, and R. Stelzer (2011) Limit theory for the largest eigenvalues of sample covariance matrices with heavy-tails, *Arxiv preprint arXiv:1108.5464*.

147 Z. Burda, J. Jurkiewicz, and M. A. Nowak (2003) Is econophysics a solid science?. *Acta Physica Polonica B* **34**, 87.

148 Z. Burda, R. Janik, J. Jurkiewicz, *et al.* (2002) Free random lévy matrices. *Physical Review E* **65**(2), 021106.

149 V. Plerou, P. Gopikrishnan, B. Rosenow, *et al.* (2002) Random matrix approach to cross correlations in financial data. *Physical Review E* **65**(6), 066126.

150 A. Utsugi, K. Ino, and M. Oshikawa (2004) Random matrix theory analysis of cross correlations in financial markets. *Physical Review E* **70**(2), 026110.

151 M. Potters, J.-P. Bouchaud, and L. Laloux (2005) Financial applications of random matrix theory: Old laces and new pieces. *Acta Physica Polonica B* **36**, 2767.

152 G. Livan and L. Rebecchi (2012) Asymmetric correlation matrices: an analysis of financial data. *The European Physical Journal B* **85**(6), 1–11.

153 G. Akemann, J. Fischmann, and P. Vivo (2010) Universal corrections and power-law tails in financial covariance matrices. *Physica A: Statistical Mechanics and its Applications* **389**, 2566–2579.

154 R. Schäfer, N. F. Nilsson, and T. Guhr (2010) Power mapping with dynamical adjustment for improved portfolio optimization. *Quantitative Finance* **10**(1), 107–119.

155 R. Schäfer and T. H. Seligman (2013) Emerging spectra of singular correlation matrices under small power-map deformations, *arXiv preprint arXiv:1304.4982*.

156 G. Akemann and P. Vivo (2008) Power law deformation of wishart–laguerre ensembles of random matrices. *Journal of Statistical Mechanics: Theory and Experiment* **2008**(09), P09002.

157 J. Fan, F. Han, and H. Liu (2013) Challenges of big data analysis, *arXiv preprint arXiv:1308.1479*.

158 G. I. Allen and P. O. Perry, "Singular Value Decomposition and High Dimensional Data," *Encyclopedia of Environmetrics*. 2013.

159 T. Palpanas (2013) Real-time data analytics in sensor networks," in *Managing and Mining Sensor Data*, 173–210, Springer.

160 Z. Bai, D. Jiang, J.-F. Yao, and S. Zheng (2009) Corrections to lrt on large-dimensional covariance matrix by rmt. *The Annals of Statistics*, 3822–3840.

161 E. Wigner (1967) Random matrices in physics. *Siam Review* **9**(1), 1–23.

162 E. Wigner (1965) Distribution laws for the roots of a random hermitian matrix. *Statistical Theories of Spectra: Fluctuations*, 446–461.

163 Z. Bai and J. Silverstein (2010) *Spectral Analysis of Large Dimensional Random Matrices*. Springer Verlag.

164 A. Edelman, B. D. Sutton, and Y. Wang (2014) Random matrix theory, numerical computation and applications. *Modern Aspects of Random Matrix Theory*, 2014, 72: 53.

165 C. Tracy and H. Widom (1994) Level-spacing distributions and the airy kernel. *Communications in Mathematical Physics* **159**(1), 151–174.

166 S. Kritchman and B. Nadler (2009) Non-parametric detection of the number of signals: hypothesis testing and random matrix theory. *Signal Processing, IEEE Transactions on* **57**(10), 3930–3941.

167 N. I. Akhiezer and N. Kemmer (1965) *The classical moment problem: and some related questions in analysis* **5**. Oliver & Boyd Edinburgh.

168 P. Lax (2002) *Functional Analysis*. John Wiley & Sons, Ltd.

169 F. Hiai and D. Petz (2000) *The Semicircle Law, Free Random Variables, and Entropy*. American Mathematical Society.

170 J. S. Geronimo and T. P. Hill (2003) Necessary and sufficient condition that the limit of stieltjes transforms is a stieltjes transform. *Journal of Approximation Theory* **121**(1), 54–60.

171 W. Rudin (1964) *Principles of mathematical analysis* **3**. McGraw-Hill New York.

172 V. Marchenko and L. Pastur (1967) Distributions of eigenvalues for some sets of random matrices. *Math. USSR-Sbornik* **1**, 457–483.

173 K. Wachter (1978) The strong limits of random matrix spectra for sample matrices of independent elements. *The Annals of Probability*, pp. 1–18.

174 J. W. Silverstein and Z. Bai (1995) On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate analysis* **54**(2), 175–192.

175 J. Silverstein (1995) Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *Journal of Multivariate Analysis* **55**(2), 331–339.

176 Z. Bai (1999) Methodologies in spectral analysis of large-dimensional random matrices, a review. *Statist. Sinica* **9**(3), 611–677.

177 I. Johnstone (2001) On the distribution of the largest eigenvalue in principal components analysis. *The Annals of statistics* **29**(2), 295–327.

178 N. El Karoui (2008) Spectrum estimation for large dimensional covariance matrices using random matrix theory. *The Annals of Statistics* **36**(6), 2757–2790.

179 P. Billingsley (2008) *Probability and measure*. John Wiley & Sons.

180 D. Jonsson (1982) Some limit theorems for the eigenvalues of a sample covariance matrix. *Journal of Multivariate Analysis* **12**(1), 1–38.

181 A. Lytova and L. Pastur (2009) Central limit theorem for linear eigenvalue statistics of random matrices with independent entries. *The Annals of Probability* **37**(5), 1778–1840.

182 T. Tao and V. Vu (2012) Random matrices: Sharp concentration of eigenvalues, *Arxiv preprint arXiv:1201.4789*.

183 A. Edelman and Y. Wang (2013) Random matrix theory and its innovative applications," in *Advances in Applied Mathematics, Modeling, and Computational Science*, 91–116, Springer.

184 B. Spain and M. G. Smith (1970) *Functions of mathematical physics*. Van Nostrand Reinhold London.

185 J. K. Hunter and B. Nachtergaele (2001) *Applied analysis*. World Scientific.

186 L. Erdős, H.-T. Yau, and J. Yin (2012) Rigidity of eigenvalues of generalized wigner matrices. *Advances in Mathematics* **229**(3), 1435–1515.

187 J. Najim (2013) Gaussian fluctuations for linear spectral statistics of large random covariance matrices, *arXiv preprint arXiv:1309.3728*.

188 Z. D. Bai and J. W. Silverstein (2004) Clt for linear spectral statistics of large-dimensional sample covariance matrices. *The Annals of Probability* **32**(1A), 553–605.

189 S. Zheng (2012) Central limit theorems for linear spectral statistics of large dimensional f-matrices, in *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **48**, 444–476, Institut Henri Poincaré.

190 Z. Bai and J. Silverstein (2004) Clt for linear spectral statistics of large-dimensional sample covariance matrices. *The Annals of Probability* **32**(1A), 553–605.

191 S. Zheng, Z. Bai, and J. Yao (2013) Clt for linear spectral statistics of random matrix $\mathbf{st}^{-1}$, *arXiv preprint arXiv:1305.1376*.

192 C. R. Rao (1973) *Linear statistical inference and its applications* **22**. John Wiley & Sons.

193 K. Wang (2013) *Optimal upper bound for the infinity norm of eigenvectors of random matrices*. PhD thesis, Rutgers, The State University of New Jersey.

194 L. Erdős, B. Schlein, and H. Yau (2009) Semicircle law on short scales and delocalization of eigenvectors for wigner random matrices. *The Annals of Probability* **37**(3), 815–852.

195 T. Tao and V. Vu (2010) Random matrices: The distribution of the smallest singular values. *Geometric And Functional Analysis* **20**(1), 260–297.

196 L. Gross (1993) Logarithmic sobolev inequalities and contractivity properties of semigroups. *Dirichlet forms*, 54–88.

197 S. Bobkov and F. Götze (1999) Exponential integrability and transportation cost related to logarithmic sobolev inequalities. *Journal of Functional Analysis* **163**(1), 1–28.

198 M. Ledoux (1999) Concentration of measure and logarithmic sobolev inequalities. *Seminaire de probabilites XXXIII*, 120–216.

199 A. Guionnet and O. Zeitouni (2000) Concentration of the spectral measure for large matrices. *Electron. Comm. Probab* **5**, 119–136.

200 E. G. Larsson, F. Tufvesson, O. Edfors, and T. L. Marzetta (2013) Massive mimo for next generation wireless systems, *arXiv preprint arXiv:1304.6690*.

201 D. Passemier (2012) *Inférence statistique dans un modèle à variances isolées de grande dimension*. PhD thesis, Université Rennes 1.

202 L. Erdos (2012) Universality for random matrices and log-gases, *arXiv preprint arXiv:1212.0839*.

203 P. Wong (2013) *Local semicircle laws for the Gaussian β-ensembles*. PhD thesis.

204 L. Erdős and H.-T. Yau (2012) Universality of local spectral statistics of random matrices. *Bulletin of the American Mathematical Society* **49**(3), 377–414.

205 L. Li and A. Soshnikov (2013) Central limit theorem for linear statistics of eigenvalues of band random matrices, *arXiv preprint arXiv:1304.6744*.

206 A. Knowles and J. Yin (2012) The outliers of a deformed wigner matrix, *arXiv preprint arXiv:1207.5619*.

207 M. S. Pinsker (1960) Information and information stability of random variables and processes. Holden-Day, 1964.

208 S. Verdu (1986) Capacity region of gaussian cdma channels: The symbolsynchronous case, in *Proc. 24th Allerton Conf*, 1025–1034.

209 G. Foschini (1996) Layered Space-Time Architecture for Wireless Communication in Fading Environment. *Bell Labs Technical Journal* **1**(2), 41–59.

210 E. Telatar (1995) *Capacity of Multiple-Antenna Gaussian Channels*. AT&T Bell Labs Internal Tech. Memo, June.

211 L. Brandenburg and A. Wyner (1974) Capacity of the gaussian channel with memory: The multivariate case. *Bell System Technical Journal* **53**(5), 745–778.

212 B. Tsybakov (1965) Transmission capacity of memoryless gaussian vector channels. *Russion*, Probl. Peredach. Inform, **1**, 26–40.

213 Y. Chen, Y. C. Eldar, and A. Goldsmith (2013) Minimax capacity loss under sub-nyquist universal sampling, *arXiv preprint arXiv:1304.7751*.

214 M. Shcherbina (2011) Central limit theorem for linear eigenvalue statistics of the wigner and sample covariance random matrices, *Arxiv preprint arXiv:1101.3249*.

215 S. ORourke and A. Soshnikov (2013) Partial linear eigenvalue statistics for wigner and sample covariance random matrices. *Journal of Theoretical Probability*, 1–19.

216 S. ORourke (2012) A note on the marchenko-pastur law for a class of random matrices with dependent entries. *Electronic Communications in Probability* **17**, 1–13.

217 A. K. Gupta and D. K. Nagar (2000) *Matrix variate distributions* **104**. CRC Press.

218 Z. Bai and J. Yao (2011) On sample eigenvalues in a generalized spiked population model. *Journal of Multivariate Analysis*, 2012, **106**(1): 167–177.

219 V. Marčenko and L. Pastur (1967) Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik* **1**, 457.

220 W. Li, J. Chen, Y. Qin, J. Yao, and Z. Bai (2013) Estimation of the population spectral distribution from a large dimensional sample covariance matrix, *arXiv preprint arXiv:1302.0355*.

221 Y. Yin (1986) Limiting spectral distribution for a class of random matrices. *Journal of multivariate analysis* **20**(1), 50–68.

222 J. Silverstein and S. Choi (1995) Analysis of the limiting spectral distribution of large dimensional random matrices. *Journal of Multivariate Analysis*, **54**(2), 295–309.

223 M. Rosenblatt (1956) Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 832–837.

224 E. Parzen (1962) On estimation of a probability density function and mode. *The annals of mathematical statistics* **33**(3), 1065–1076.

225 P. Hall (1984) An optimal property of kernel estimators of a probability density. *Journal of the Royal Statistical Society. Series B (Methodological)*, 134–138.

226 B. W. Silverman (1986) *Density estimation for statistics and data analysis* **26**. CRC Press.

227 L. Devroye and G. Lugosi (2001) *Combinatorial methods in density estimation*. Springer.

228 B.-Y. Jing, G. Pan, Q.-M. Shao, and W. Zhou (2010) Nonparametric estimate of spectral density functions of sample covariance matrices: A first step. *The Annals of Statistics* **38**(6), 3724–3750.

229 G. Pan, Q.-M. Shao, and W. Zhou (2010) Central limit theorem of nonparametric estimate of spectral density functions of sample covariance matrices, *arXiv preprint arXiv:1008.3954.*

230 Raginsky, M. (2011) *Concentration inequalities*, http://maxim.ece.illinois.edu/teaching/spring11/notes/concentration.pdf (accessed. September 11, 2016)

231 C. McDiarmid (1989) On the method of bounded differences. *Surveys in combinatorics* **141**(1), 148–188.

232 D. N. C. Tse and S. V. Hanly (1999) Linear multiuser receivers: Effective interference, effective bandwidth and user capacity. *Information Theory, IEEE Transactions on* **45**(2), 641–657.

233 Z. Bai and J. Silverstein (1998) No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *The Annals of Probability* **26**(1), 316–345.

234 G. Pan, Q. Shao, and W. Zhou (2011) Universality of sample covariance matrices: Clt of the smoothed empirical spectral distribution, *Arxiv preprint arXiv:1111.5420*.

235 W. Wu (2005) Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences of the United States of America* **102**(40), 14150.

236 W. Wu (2011) Asymptotic theory for stationary processes, *Statistics and Its Interface, 0*, 1–20.

237 M. B. Priestley (1988) Non-Linear and Non-Stationary Time Series Analysis, The Red Republican & The Friend of the people. Barnes & Noble Inc. 1989: 385–386.

238 W. B. Wu (2007) Strong invariance principles for dependent random variables. *The Annals of Probability* **35**(6), 2294–2320.

239 M. Banna, F. Merlevede, *et al.* (2013) Limiting Spectral Distribution of Large Sample Covariance Matrices Associated with a Class of Stationary Processes. *Journal of Theoretical Probability*, 2015, **28**(2): 745–783

240 M. Forni, M. Hallin, M. Lippi, and L. Reichlin (2005) The generalized dynamic factor model. *Journal of the American Statistical Association* **100**(471).

241 R. Vautard, P. Yiou, and M. Ghil (1992) Singular-spectrum analysis: A toolkit for short, noisy chaotic signals. *Physica D: Nonlinear Phenomena* **58**(1), 95–126.

242 A. Zhigljavsky (2012) Singular spectrum analysis: Present, past and future., in *International Conference on Forecasting Economic and Financial Systems*, (Beijing, China).

243 Z. Bai and J. W. Silverstein (2012) No eigenvalues outside the support of the limiting spectral distribution of information-plus-noise type matrices. *Random Matrices: Theory and Applications* 2012, **01**(1): 1150004.

244 B. Jin, C. Wang, Z. Bai, *et al.* A note on the limiting spectral distribution of a symmetrized auto-cross covariance matrix. Annals of Applied Probability, 2014, **24**(3): 333–340.

245 H. Liu, A. Aue, and D. Paul (2013) On the marcenko-pastur law for linear time series, *arXiv preprint arXiv:1310.7270.*

246 A. Auffinger, G. Ben Arous, and S. Péché (2009) Poisson convergence for the largest eigenvalues of heavy tailed random matrices. *Ann. Inst. Henri Poincaré Probab. Stat* **45**(3), 589–610.

247 N. Xia (2013) *LIMITING BEHAVIOR OF EIGENVECTORS OF LARGE DIMENSIONAL RANDOM MATRICES*. Phd dissertation, National University of Singapore.

248 D. Passemier and J.-F. Yao (2013) Variance estimation and goodness-of-fit test in a high-dimensional strict factor model, *arXiv preprint arXiv:1308.3890*.

249 G. Pan (2011) Comparison between two types of large sample covariance matrices. Annales De L Institut Henri Poincaré Probabilités Et Statistiques, 2014, **50**(2): 655–677.

250 C. Wang, B. Jin, and B. Miao (2011) On limiting spectral distribution of large sample covariance matrices by varma (p, q). *Journal of Time Series Analysis* **32**(5), 539–546.

251 J. Yao (2012) A note on a marčenko–pastur type theorem for time series. *Statistics & Probability Letters* **82**(1), 22–28.

252 O. Pfaffel and E. Schlemm (2012) Eigenvalue distribution of large sample covariance matrices of linear processes, *arXiv preprint arXiv:1201.3828*.

253 O. Pfaffel (2012) Eigenvalues of large random matrices with dependent entries and strong solutions of sdes. *Doctor Thesis (Technische Universitat Munchen, Lehrstuhl fur Mathematische Statistik, 2013)*.

254 A. Chakrabarty, R. S. Hazra, and P. Roy (2013) Maximum eigenvalue of symmetric random matrices with dependent heavy tailed entries, *arXiv preprint arXiv:1309.1407*.

255 C. W. Granger (2001) Macroeconometrics–past and future. *Journal of Econometrics* **100**(1), 17–19.

256 A. Clauset, C. R. Shalizi, and M. E. Newman (2009) Power-law distributions in empirical data. *SIAM review* **51**(4), 661–703.

257 L. Debnath and P. Mikusiński (2005) *Hilbert spaces with applications*. Academic press.

258 J.-P. Bouchaud and M. Potters (2000) *Theory of financial risks: from statistical physics to risk management* **12**. Cambridge University Press Cambridge.

259 A. Nica and R. Speicher (2006) Lectures on the combinatorics of free probability, London mathematical society. Cambridge UK, 2010.

260 O. E. Barndorff-Nielsen and S. Thorbjornsen (2002) Lévy laws in free probability. *Proceedings of the National Academy of Sciences* **99**(26), 16568–16575.

261 Janik R. A., Nowak M. A., Papp G, et al. Various Shades of Blue's Functions. Acta Physica Polonica, 1997, 28(12).

262 R. Müller (2003) Applications of large random matrices in communications engineering, in *Proc. Int. Conf. on Advances Internet, Process., Syst., Interdisciplinary Research (IPSI), Sveti Stefan, Montenegro*.

263 W. Hachem, P. Loubaton, and J. Najim (2011) Applications of large random matrices to digital communications and statistical signal processing. EUSIPCO, September. Presentation (133 slides).

264 L. Pastur (2005) A simple approach to the global regime of gaussian ensembles of random matrices. *Ukrainian Mathematical Journal* **57**(6), 936–966.

265 W. Hachem, P. Loubaton, and J. Najim (2007) Deterministic equivalents for certain functionals of large random matrices, *The Annals of Applied Probability* **17**(3), 875–930.

266 P. Vallet (2011) Random matrix theory and applications to statistical signal processing." PhD Dissertation, November. Universite Paris-Est.

267 G. Pan (2010) Strong convergence of the empirical distribution of eigenvalues of sample covariance matrices with a perturbation matrix. *Journal of Multivariate Analysis* **101**(6), 1330–1338.

268 A. Gittens and J. Tropp (2011) Tail bounds for all eigenvalues of a sum of random matrices, *Arxiv preprint arXiv:1104.4513*.

269 D. Voiculescu (1987) Multiplication of certain non-commuting random variables. *Journal of Operator Theory* **18**(2), 223–235.

270 D. Voiculescu (1991) Limit laws for random matrices and free products. *Inventiones mathematicae* **104**(1), 201–220.

271 R. Muller, D. Guo, and A. Moustakas (2008) Vector precoding for wireless mimo systems and its replica analysis. *Selected Areas in Communications, IEEE Journal on* **26**(3), 530–540.

272 N. R. Rao and R. Speicher (2007) Multiplication of free random variables and the s-transform: the case of vanishing mean. *Electronic Communications in Probability* **12**, 248–258.

273 J. M. Lindsay and V. Pata (1997) Some weak laws of large numbers in noncommutative probability. *Mathematische Zeitschrift* **226**(4), 533–543.

274 U. Haagerup and S. Möller (2013) The law of large numbers for the free multiplicative convolution. *Operator Algebra and Dynamics*, 157–186.

275 G. Tucci and P. Whiting (2011) Eigenvalue results for large scale random vandermonde matrices with unit complex entries. *Information Theory, IEEE Transactions on* **57**(6), 3938–3954.

276 P. Billingsley (1968) *Weak Convergence of Probability Measures*. John Wiley.

277 M. Desgroseilliers, O. Lévêque, and E. Preissmann (2013) Partially random matrices in line-of-sight wireless networks. *Proc., IEEE Asilomar, Pacific Grove, CA*.

278 M. Desgroseilliers, O. Lévêque, and E. Preissmann (2013) Spatial degrees of freedom of mimo systems in line-of-sight environment, in *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, 834–838,

279 T. Kailath, A. Sayed, and B. Hassibi (2000) *Linear Estimation*. Prentice Hall.

280 M. Politi, E. Scalas, D. Fulger, and G. Germano (2010) Spectral densities of wishart-lévy free stable random matrices. *The European Physical Journal B* **73**(1), 13–22.

281 O. Ryan and M. Debbah (2009) Asymptotic behavior of random vandermonde matrices with entries on the unit circle. *Information Theory, IEEE Transactions on* **55**(7), 3115–3147.

282 R. R. Müller, G. Alfano, B. M. Zaidel, and R. de Miguel (2013) Applications of large random matrices in communications engineering, *arXiv preprint arXiv:1310.5479*.

283 G. H. Tucci and P. A. Whiting (2012) Asymptotic behavior of the maximum and minimum singular value of random vandermonde matrices. *Journal of Theoretical Probability*, 1–37.

284 H. S. Dhillon and G. Caire (2014) Scalability of line-of-sight massive mimo mesh networks for wireless backhaul, in *submitted to IEEE Intl. Symposium on Information Theory, Honolulu, HI*.

285 H. S. Dhillon and G. Caire, Information theoretic upper bound on the capacity of wireless backhaul networks. 2014: 251–255.

286 Y. Chen, A. J. Goldsmith, and Y. C. Eldar (2013) Non-Asymptotic Analysis of Random Vector Channels. 2013.

287 J. R. Ipsen and M. Kieburg (2013) Weak commutation relations and eigenvalue statistics for products of rectangular random matrices, *arXiv preprint arXiv:1310.4154*.

288 R. Remmert (1991) *Theory of complex functions* **122**. Springer.

289 Z. Burda, R. Janik, and M. Nowak (2011) Multiplication law and s transform for non-hermitian random matrices. *Physical Review E* **84**(6), 061125.

290 K. J. Dykema (2006) On the s-transform over a banach algebra. *Journal of Functional Analysis* **231**(1), 90–110.

291 Biane, P. and Lehner, F. (2001) Computation of some examples of Brown's spectral measure in free probability. Mathematics, **2001**(2): 181–211.

292 U. Haagerup and F. Larsen (2000) Brown's spectral distribution measure for r-diagonal elements in finite von neumann algebras. *Journal of Functional Analysis* **176**(2), 331–367.

293 G. H. Tucci (2010) Limits laws for geometric means of free random variables. *Indiana University mathematics journal* **59**(1), 1–13.

294 D. Voiculescu (2000) Lectures on free probability theory. *Lectures on probability theory and statistics (Saint-Flour, 1998)* **1738**, 279–349.

295 V. Kargin (2008) On asymptotic growth of the support of free multiplicative convolutions. *Electronic Communications in Probability* **13**, 415–412.

296 O. Arizmendi and C. Vargas (2012) Products of free random variables and k-divisible partitions, *arXiv preprint arXiv:1201.5825*.

297 N. J. Higham (2008) *Functions of matrices: theory and computation*. Siam.

298 K. Knopp (1957) *Theory and application of infinite series*. Blackie Son.

299 M. Krishnapur (2009) From random matrices to random analytic functions. *The Annals of Probability* **37**(1), 314–346.

300 T. Rogers (2010) Universal sum and product rules for random matrices. *Journal of Mathematical Physics* **51**, 093304.

301 P. J. Forrester and A. Mays (2012) Pfaffian point process for the gaussian real generalised eigenvalue problem. *Probability Theory and Related Fields* **154**(1–2), 1–47.

302 C. Bordenave (2011) On the spectrum of sum and product of non-hermitian random matrices. *Electronic Communications in Probability* **16**, 104–113.

303 E. E. T. Whittaker and G. Watson (1980) *A course of modern analysis*. Cambridge University Press.

304 T. J. I. Bromwich (1991) *An Introduction to the Theory of Infinite Series*. 1951, **78**(2020): 242–242.

305 N. Alexeev, F. Götze, and A. Tikhomirov (2010) Asymptotic distribution of singular values of powers of random matrices. *Lithuanian mathematical journal* **50**(2), 121–132.

306 T. Banica, S. Belinschi, M. Capitaine, and B. Collins (2011) Free bessel laws. *Canad. J. Math* **63**(1), 3–37.

307 F. Benaych-Georges (2010) On a surprising relation between the marchenko–pastur law, rectangular and square free convolutions, in *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **46**, 644–652, Institut Henri Poincaré.

308 H. Bateman and A. Erdélyi (1981) *Higher transcendental functions*. Krieger.

309 K. Życzkowski, K. A. Penson, I. Nechita, and B. Collins (2011) Generating random density matrices. *Journal of Mathematical Physics* **52**, 062201.

310 T. Neuschel (2013) Plancherel-rotach formulae for average characteristic polynomials of products of ginibre random matrices and the fuss-catalan distribution. *Random Matrices Theory & Applications*, 2013, **3**(01):14500031-145000318.

311 Z. Burda, M. Nowak, and A. Swiech (2012) New spectral relations between products and powers of isotropic random matrices, *arXiv preprint arXiv:1205.1625*.

312 Z. Burda, G. Livan, and A. Swiech (2013) Commutative law for products of infinitely large isotropic random matrices, *arXiv preprint arXiv:1303.5360*.

313 V. I. Oseledec (1968) A multiplicative ergodic theorem. lyapunov characteristic numbers for dynamical systems. *Trans. Moscow Math. Soc* **19**(2), 197–231.

314 C. M. Newman (1986) The distribution of lyapunov exponents: Exact results for random matrices. *Communications in mathematical physics* **103**(1), 121–126.

315 P. J. Forrester (2013) Lyapunov exponents for products of complex gaussian random matrices. *Journal of Statistical Physics*, 1–13.

316 D. Mannion (1993) Products of 2 × 2 random matrices. *The Annals of Applied Probability* **3**(4), 1189–1218.

317 J. Marklof, Y. Tourigny, and L. Wolowski (2008) Explicit invariant measures for products of random matrices. *Transactions of the American Mathematical Society* **360**(7), 3391–3427.

318 M. Pollicott (2010) Maximal lyapunov exponents for random matrix products. *Inventiones mathematicae* **181**(1), 209–226.

319 V. Kargin (2013) On the largest lyapunov exponent for products of gaussian matrices, *arXiv preprint arXiv:1306.6576*.

320 S. Skipetrov and A. Goetschy (2011) Eigenvalue distributions of large euclidean random matrices for waves in random media. *Journal of Physics A: Mathematical and Theoretical* **44**, 065102.

321 A. Goetschy and S. Skipetrov (2011) Non-hermitian euclidean random matrix theory. *PHYSICAL REVIEW E Phys Rev E* **81**, 011150. American Physical Society.

322 A. Jarosz and M. A. Nowak (2004) A novel approach to non-hermitian random matrix models, *arXiv preprint math-ph/0402057*.

323 A. Jarosz and M. A. Nowak (2006) Random hermitian versus random non-hermitian operators unexpected links. *Journal of Physics A: Mathematical and General* **39**(32), 10107.

324 M. Mehta (1967) *Random matrices and the statistical theory of energy levels*. Academic Press.

325 G. Pan and W. Zhou (2010) Circular law, extreme singular values and potential theory. *Journal of Multivariate Analysis* **101**(3), 645–656.

326 T. Tao and V. Vu (2011) Random matrices: Universality of local eigenvalue statistics. *Acta mathematica*, 1–78.

327 T. Tao and V. Vu (2007) Random matrices: the circular law, *Arxiv preprint arXiv:0708.2895*.

328 C. Bordenave and D. Chafaï (2012) Around the circular law. *Probability Surveys* **9**.

329 C. Bordenave and D. Chafaı (2013) The circular law.

330 S. Coleri, M. Ergen, A. Puri, and A. Bahai (2002) Channel estimation techniques based on pilot arrangement in ofdm systems. *Broadcasting, IEEE Transactions on* **48**(3), 223–229.

331 J.-J. Van de Beek, O. Edfors, M. Sandell, *et al.* (1995) On channel estimation in ofdm systems, in *Vehicular Technology Conference, 1995 IEEE 45th*, vol. 2, 815–819.

332 O. Edfors, M. Sandell, J.-J. Van de Beek, *et al.* (1998) Ofdm channel estimation by singular value decomposition. *Communications, IEEE Transactions on* **46**(7), 931–939.

333 T. Tao (2011) Outliers in the spectrum of iid matrices with bounded rank perturbations. *Probability Theory and Related Fields*, 1–33.

334 C. Bordenave, P. Caputo, and D. Chafai (2013) Spectrum of markov generators on sparse random graphs, *arXiv:1202.0644v2*, p. 33, March.

335 J. Baik, G. Ben Arous, and S. Péché (2005) Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability* **33**(5), 1643–1697.

336 A. Guionnet and O. Zeitouni (2012) Support convergence in the single ring theorem. *Probability Theory and Related Fields* **154**(3–4), 661–675.

337 M. Rudelson and R. Vershynin (2013) Invertibility of random matrices: unitary and orthogonal perturbations. *Journal of the American Mathematical Society*.

338 H. Sommers, A. Crisanti, H. Sompolinsky, and Y. Stein (1988) Spectrum of large random asymmetric matrices. *Physical review letters* **60**(19), 1895–1898.

339 A. Naumov (2012) The elliptic law, *arXiv preprint arXiv:1201.1639*.

340 A. Naumov (2012) Universality of some models of random matrices and random processes. ARCHIVE Proceedings of the Institution of Mechanical Engineers Part J *Journal of Engineering Tribology* 1994–1996 (vols 208–210), 2007, **221**(2): 161–164.

341 H. Nguyen and S. O'Rourke (2012) The elliptic law, *arXiv preprint arXiv:1208.5883*.

342 M. Capitaine, C. Donati-Martin, and D. Féral (2009) The largest eigenvalues of finite rank deformation of large wigner matrices: convergence and nonuniversality of the fluctuations. *The Annals of Probability* **37**(1), 1–47.

343 M. Capitaine, C. Donati-Martin, D. Féral, and M. Février (2011) Free convolution with a semi-circular distribution and eigenvalues of spiked deformations of wigner matrices. preprint (2010). *Elec. J. Probab.* **16**(64), 1750–1792.

344 Z. Füredi and J. Komlós (1981) The eigenvalues of random symmetric matrices. *Combinatorica* **1**(3), 233–241.

345 T. Tao (2013) Outliers in the spectrum of iid matrices with bounded rank perturbations. *Probability Theory and Related Fields* **155**(1–2), 231–263.

346 R. R. Müller, M. Vehkaperä, and L. Cottatellucci (2013) Blind Pilot Decontamination. International ITG Workshop on Smart Antennas. VDE, 2013: 1–6.

347 R. R. Müller, M. Vehkaperä, and L. Cottatellucci (2013) Analysis of blind pilot decontamination, in *Proceedings of the 47th Annual Asilomar Conference on Signals, Systems, and Computers*.

348 B. Cakmak, R. R. Müller, and B. H. Fleury (2013) Beyond multiplexing gain in large mimo systems, *arXiv preprint arXiv:1306.2595*.

349 R. R. Muller and B. Cakmak (2012) Channel modelling of mu-mimo systems by quaternionic free probability, in *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*, 2656–2660.

350 B. Çakmak, R. R. Müller, and B. H. Fleury (2013) Beyond multiplexing gain in large mimo systems, *arXiv preprint arXiv:1306.2595*.

351 Z. Burda, A. Jarosz, G. Livan, (2011) Eigenvalues and singular values of products of rectangular gaussian random matrices (the extended version), *Arxiv preprint arXiv:1103.3964*.

352 G. Akemann and Z. Burda (2012) Universal microscopic correlation functions for products of independent ginibre matrices. *Journal of Physics A: Mathematical and Theoretical* **45**(46), 465201.

353 G. Akemann, M. Kieburg, and L. Wei (2013) Singular value correlation functions for products of wishart random matrices. *Journal of Physics A: Mathematical and Theoretical* **46**(27), 275205.

354 G. Akemann, J. R. Ipsen, and M. Kieburg (2013) Products of rectangular random matrices: Singular values and progressive scattering. *Physical Review E* **88**(5), 052118.

355 P. J. Forrester (2013) Probability of all eigenvalues real for products of standard gaussian matrices, *arXiv preprint arXiv:1309.7736*.

356 V. Y. Protasov and R. Jungers (2013) Lower and upper bounds for the largest lyapunov exponent of matrices. *Linear Algebra and its Applications*.

357 V. Kargin (2008) Lyapunov exponents of free operators. *Journal of Functional Analysis* **255**(8), 1874–1888.

358 A. Goetschy and S. Skipetrov (2013) Euclidean random matrices and their applications in physics, *arXiv preprint arXiv:1303.2880*.

359 A. Goetschy (2011) *Lumière dans les milieux atomiques désordonnés: théorie des matrices euclidiennes et lasers aléatoires*. PhD thesis, Université de Grenoble.

360 A. Goetschy and S. Skipetrov (2011) Euclidean matrix theory of random lasing in a cloud of cold atoms. *EPL (Europhysics Letters)* **96**(3), 34005.

361 M.-T. Rouabah, M. Samoylova, R. Bachelard, *et al.* (2014) Coherence effects in scattering order expansion of light by atomic clouds, *arXiv preprint arXiv:1401.5704*.

362 C. Bordenave, P. Caputo, and D. Chafaï (2008) Circular law theorem for random markov matrices. *Probability Theory and Related Fields*, 1–29.

363 C. Bordenave (2013) On euclidean random matrices in high dimension. *Electronic Communications in Probability* **18**, 1–8.

364 T. Jiang (2013) Distributions of eigenvalues of large euclidean matrices generated from lp balls and spheres. *Linear Algebra and its Applications*.

365 Y. Do and V. Vu (2013) The spectrum of random kernel matrices: universality results for rough and varying kernels. *Random Matrices Theory & Applications*, **2**(03).

366 X. Cheng (2013) Random Matrices in High-dimensional Data Analysis. Princeton NJ Princeton University, 2013.

367 F. Benaych-Georges and R. Nadakuditi (2012) The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis* **111**, 120–135.

368 S. O'Rourke and D. Renfrew (2013) Low rank perturbations of large elliptic random matrices, *arXiv preprint arXiv:1309.5326*.

369 C. Stein (1975) Estimation of a covariance matrix. reitz lecture, in *Reitz Lecture, IMS-ASA Annual Meeting*. (Also unpublished lecture notes.)

370 A. P. Dempster (1972) Covariance selection. *Biometrics*, 157–175.

371 T. Anderson (2003) An introduction to multivariate statistical analysis. *Wiley series in probability and mathematical statistics*. Wiley, 1984.

372 W. James and C. Stein (1961) Estimation with quadratic loss, in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability* **1**, 361–379.

373 C. Stein (1956) Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, in *Proceedings of the Third Berkeley symposium on mathematical statistics and probability* **1**, 197–206.

374 D. I. Warton (2008) Penalized normal likelihood and ridge regularization of correlation and covariance matrices. *Journal of the American Statistical Association* **103**(481).

375 O. Ledoit and M. Wolf (2004) A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis* **88**(2), 365–411.

376 Y. Sheena and A. K. Gupta (2003) Estimation of the multivariate normal covariance matrix under some restrictions. *Statistics & Decisions/International mathematical Journal for stochastic methods and models* **21**(4/2003), 327–342.

377 S. Boyd and L. Vandenberghe (2004) *Convex optimization*. Cambridge Univ Pr.

378 M. Grant, S. Boyd, and Y. Ye (2008) Cvx: Matlab software for disciplined convex programming.

379 M. Andersen, J. Dahl, and L. Vandenberghe, CVXOPT: Python software for convex optimization. 2009.

380 R. Bellman, R. E. Bellman, R. E. Bellman, and R. E. Bellman (1970) *Introduction to matrix analysis* **10**. SIAM.

381 X. Deng and K.-W. Tsui (2013) Penalized covariance matrix estimation using a matrix-logarithm transformation. *Journal of Computational and Graphical Statistics* **22**(2), 494–512.

382 C. M. Stein (1981) Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, 1135–1151.

383 I. Johnstone (2007) *Gaussian estimation: Sequence and wavelet models*. Springer Texts in Statistics, Manuscript, December.

384 E. J. Candes, C. A. Sing-Long, and J. D. Trzasko (2012) Unbiased risk estimates for singular value thresholding and spectral estimators, *arXiv preprint arXiv:1210.4139*.

385 E. Candes, C. Sing-Long, and J. Trzasko, Unbiased Risk Estimates for Singular Value Thresholding and Spectral Estimators. *IEEE Transactions on Signal Processing*, 2013, **61**(19): 4643–4657.

386 C.-A. Deledalle, S. Vaiter, G. Peyré, *et al.* (2012) Risk estimation for matrix recovery with spectral regularization, *arXiv preprint arXiv:1205.1482*.

387 S. Oymak and B. Hassibi (2013) On a relation between the minimax risk and the phase transitions of compressed recovery. Communication, Control, and Computing. 2012: 1018–1025.

388 S. Oymak and B. Hassibi (2013) Asymptotically exact denoising in relation to compressed sensing, *arXiv preprint arXiv:1305.2714*.

389 D. L. Donoho and M. Gavish (2013) Minimax risk of matrix denoising by singular value thresholding, *arXiv preprint arXiv:1304.2085*.

390 M. Verbanck, J. Josse, and F. Husson (2013) Regularised pca to denoise and visualise data, *arXiv preprint arXiv:1301.4649*.

391 A. A. Shabalin and A. B. Nobel (2013) Reconstruction of a low-rank matrix in the presence of gaussian noise. *Journal of Multivariate Analysis*.

392 J. Baik and J. Silverstein (2006) Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis* **97**(6), 1382–1408.

393 D. Paul (2007) Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica* **17**(4), 1617.

394 B. Nadler (2008) Finite sample approximation results for principal component analysis: A matrix perturbation approach. *The Annals of Statistics* **36**(6), 2791–2817.

395 S. Lee, F. Zou, and F. Wright (2010) Convergence and prediction of principal component scores in high-dimensional settings. *Annals of statistics* **38**(6), 3605.

396 L. Györfi, I. Vajda, and E. Van Der Meulen (1996) Minimum kolmogorov distance estimates of parameters and parametrized distributions. *Metrika* **43**(1), 237–255.

397 G. Golub and W. Kahan (1965) Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial & Applied Mathematics, Series B: Numerical Analysis* **2**(2), 205–224.

398 D. L. Donoho and M. Gavish (2013) The optimal hard threshold for singular values is 4/sqrt (3), *arXiv preprint arXiv:1305.5870*.

399 C. Eckart and G. Young (1936) The approximation of one matrix by another of lower rank. *Psychometrika* **1**(3), 211–218.

400 L. Mirsky (1960) Symmetric gauge functions and unitarily invariant norms. *The quarterly journal of mathematics* **11**(1), 50–59.

401 L. M. Le Cam (1960) *Locally asymptotically normal families of distributions: certain approximations to families of distributions and their use in the theory of estimation and testing hypotheses* **3**. University of California Press.

402 O. Ledoit and M. Wolf (2004) A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* **88**(2), 365–411.

403 O. Ledoit and M. Wolf (2003) Honey, I shrunk the sample covariance matrix. *UPF Economics and Business Working Paper*, no. 691.

404 B. Efron (1982) Maximum likelihood and decision theory. *The annals of Statistics*, 340–356.

405 O. Ledoit and M. Wolf (2003) Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance* **10**(5), 603–621.

406 J. Schäfer and K. Strimmer (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology* **4**(1), 32.

407 U. Grenander and J. Silverstein (1977) Spectral analysis of networks with random topologies. *SIAM Journal on Applied Mathematics*, 499–519.

408 Y. Yin and P. Krishnaiah (1983) A limit theorem for the eigenvalues of product of two random matrices. *Journal of Multivariate Analysis* **13**(4), 489–507.

409 Z. Bai, B. Miao, and G. Pan (2007) On asymptotics of eigenvectors of large sample covariance matrix. *The Annals of Probability* **35**(4), 1532–1572.

410 G. Pan and W. Zhou (2008) Central limit theorem for signal-to-interference ratio of reduced rank linear receiver. *The Annals of Applied Probability* **18**(3), 1232–1270.

411 O. Ledoit and S. Péché (2011) Eigenvectors of some large sample covariance matrix ensembles. *Probability theory and related fields* **151**(1–2), 233–264.

412 M. Perlman (2007) Multivariate statistical analysis. Wiley and Sons, New York, NY, 1984.

413 G. Pan and W. Zhou (2011) Central limit theorem for hotellings t2 statistic under large dimension. *The Annals of Applied Probability* **21**(5), 1860–1910.

414 X. Mestre and M. Lagunas (2006) Finite sample size effect on minimum variance beamformers: Optimum diagonal loading factor for large arrays. *Signal Processing, IEEE Transactions on* **54**(1), 69–82.

415 O. Ledoit and M. Wolf (2012) Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics* **40**(2), 1024–1060.

416 C. Wang, G. Pan, and L. Cao (2012) A shrinkage estimation for large dimensional precision matrices using random matrix theory, *arXiv preprint arXiv:1211.2400*.

417 L. S. Chen, D. Paul, R. L. Prentice, and P. Wang (2011) A regularized hotellings t2 test for pathway analysis in proteomic studies. *Journal of the American Statistical Association* **106**(496), 1345–1360.

418 F. Rubio and X. Mestre (2011) Spectral convergence for a general class of random matrices. *Statistics & Probability Letters* **81**(5), 592–602.

419 T. Bodnar, A. K. Gupta, and N. Parolya (2013) Optimal linear shrinkage estimator for large dimensional precision matrix, *arXiv preprint arXiv:1308.0931*.

420 T. Bodnar, A. K. Gupta, and N. Parolya (2013) On the strong convergence of the optimal linear shrinkage estimator for large dimensional covariance matrix, *arXiv preprint arXiv:1308.2608*.

421 J. Vinogradova, R. Couillet, W. Hachem, *et al.* (2013) A new method for source detection, power estimation, and localization in large sensor networks under noise with unknown statistics, in *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing*.

422 J. Vinogradova, R. Couillet, and W. Hachem (2013) Statistical inference in large antenna arrays under unknown noise pattern, *arXiv preprint arXiv:1301.0306*.

423 S. M. Kay (1998) *Fundamentals of Statistical Signal Processing, Vol. II: Detection Theory*. PTR Prentice Hall, 1993.

424 A. Kammoun, R. Couillet, J. Najim, and M. Debbah (2013) Performance of capacity inference methods under colored interference. *IEEE Trans. Information Theory* **59**(2), 1129–1148.

425 R. R. Nadakuditi (2013) Optshrink: An algorithm for improved low-rank signal matrix denoising by optimal, data-driven singular value shrinkage, *arXiv preprint arXiv:1306.6042*.

426 N. Rao, J. Mingo, R. Speicher, and A. Edelman (2008) Statistical eigen-inference from large wishart matrices. *The Annals of Statistics* **36**(6), 2850–2885.

427 J. Chen, B. Delyon, and J.-F. Yao (2011) On a model selection problem from high-dimensional sample covariance matrices. *Journal of Multivariate Analysis* **102**(10), 1388–1398.

428 Z. Bai, J. Hu, and W. Zhou (2012) Convergence rates to the marchenko–pastur type distrbution. *Stochastic Processes and their Applications* **122**, 68–92.

429 O. Ledoit and M. Wolf (2013) Optimal estimation of a large-dimensional covariance matrix under steins loss. *University of Zurich Department of Economics Working Paper*, no. 122.

430 O. Ledoit and M. Wolf (2013) Spectrum estimation: A unified framework for covariance matrix estimation and pca in large dimensions, *Available at SSRN 2198287*.

431 K. Lounici (2012) High-dimensional covariance matrix estimation with missing observations, *Arxiv preprint arXiv:1201.2577*.

432 J. Won, J. Lim, S. Kim, and B. Rajaratnam (2009) Maximum likelihood covariance estimation with a condition number constraint, Technical Report No. 2009-10, Stanford University, Department of Statistics.

433 A. K. Gupta, T. Varga, and T. Bodnar (2013) *Elliptically contoured models in statistics and portfolio theory*. Springer, second edition ed.

434 S. Dallaporta (2012) Eigenvalue variance bounds for wigner and covariance random matrices. *Random Matrices: Theory and Applications* **01**(3), 1250007.

435 S. Dallaporta (2012) *Quelques aspects de l'étude quantitative de la fonction de comptage et des valeurs propres de matrices aléatoires*. PhD thesis, Université de Toulouse, Université Toulouse III-Paul Sabatier.

436 L. L. A. Pastur and M. V. Ŝerbina (2011) *Eigenvalue distribution of large random matrices* **171**. AMS Bookstore.

437 H. Cremér (1999) *Mathematical Methods of Statistics (PMS-9)* **9**. Princeton university press.

438 B. V. Gnedenko (2005) *The Theory of Probability: And the Elements of Statistics* **132**. AMS Bookstore.

439 W. Feller (2008) *An introduction to probability theory and its applications*, **2**. John Wiley & Sons.

440 V. V. Petrov (1995) Limit theorems of probability theory: sequences of independent random variables. *Journal of Applied Statistics* (4), 575.

441 B. B. V. Gnedenko and A. Y. Khinchin (1962) *An elementary introduction to the theory of probability*, **155**. Courier Dover Publications.

442 M. Rudelson and R. Vershynin (2010) Non-asymptotic theory of random matrices: extreme singular values, *Arxiv preprint arXiv:1003.2990*.

443 C. Villani (2003) *Topics in optimal transportation*. Ams Graduate Studies in Mathematics, 2003:370.

444 S. Chatterjee and A. Bose (2004) A new method for bounding rates of convergence of empirical spectral distributions. *Journal of Theoretical Probability* **17**(4), 1003–1019.

445 K. Davidson and S. Szarek (2001) Local operator theory, random matrices and banach spaces. *Handbook of the geometry of Banach spaces* **1**, 317–366.

446 M. Ledoux (2001) *The concentration of measure phenomenon*. Mathematical Surveys & Monographs. 2001.

447 S. G. Bobkov and C. Houdré (1997) Isoperimetric constants for product probability measures. *The Annals of Probability*, 184–205.

448 S. Bobkov and F. Götze (2010) Concentration of empirical distribution functions with applications to non-iid models. *Bernoulli* **16**(4), 1385–1414.

449 V. M. Zolotarev (1971) Estimates of the difference between distributions in the lévy metric. *Trudy Matematicheskogo Instituta im. VA Steklova* **112**, 224–231.

450 J. H. Kim (2013) *Concentration of Empirical Distribution Functions for Dependent Data under Analytic Hypotheses*. PhD thesis, University of Minnesota.

451 L. Arnold (1971) On wigner's semicircle law for the eigenvalues of random matrices. *Probability Theory and Related Fields* **19**(3), 191–198.

452 D. L. Hanson and F. T. Wright (1971) A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics* **42**(3), 1079–1083.

453 D. Hsu, S. M. Kakade, and T. Zhang (2011) A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 2011, **17**(25): 1–6.

454 V. Vu and K. Wang (2013) Random weighted projections, random quadratic forms and random eigenvectors, *arXiv preprint arXiv:1306.3099*.

455 H. Nguyen and V. Vu (2011) Random matrices: Law of the determinant, *Arxiv preprint arXiv:1112.0752*.

456 L. Erdos and B. Farrell (2012) Local eigenvalue density for general manova matrices, *arXiv preprint arXiv:1207.0031*.

457 K. W. Wachter (1980) The limiting empirical measure of multiple discriminant ratios. *The Annals of Statistics*, 937–957.

458 R. Schmidt (1986) Multiple emitter location and signal parameter estimation. *Antennas and Propagation, IEEE Transactions on* **34**(3), 276–280.

459 M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul (1998) *An introduction to variational methods for graphical models*. Springer.

460 M. E. Tipping and C. M. Bishop (1999) Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61**(3), 611–622.

461 F. Benaych-Georges (2008) On a surprising relation between rectangular and square free convolutions, *Arxiv preprint arXiv:0807.0505*.

462 F. Benaych-Georges (2005) Rectangular random matrices, related free entropy and free fisher's information, *Arxiv preprint math/0512081*.

463 F. Benaych-Georges (2009) Rectangular random matrices, related convolution. *Probability Theory and Related Fields* **144**(3), 471–515.

464 F. Benaych-Georges and R. R. Nadakuditi (2011) The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics* **227**(1), 494–521.

465 M. Peng (2012) Eigenvalues of deformed random matrices, *arXiv preprint arXiv:1205.0572*.

466 I. M. Johnstone and D. M. Titterington (2009) Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **367**(1906), 4237–4253.

467 R. R. Nadakuditi (2013) When are the most informative components for inference also the principal components?, *arXiv preprint arXiv:1302.1232*.

468 A. Buja, D. Cook, H. Hofmann, *et al.* (2009) Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **367**(1906), 4361–4383.

469 O. Ledoit and M. Wolf (2002) Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Annals of Statistics*, 1081–1102.

470 A. Onatski (2012) Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics* **168**(2), 244–258.

471 A. P. Dempster (1958) A high dimensional two sample significance test. *The Annals of Mathematical Statistics* **29**(4), 995–1010.

472 Z. Bai and H. Saranadasa (1996) Effect of high dimension: by an example of a two sample problem. *Statistica Sinica* **6**, 311–330.

473 D. Jiang, T. Jiang, and F. Yang (2012) Likelihood ratio tests for covariance matrices of high-dimensional normal distributions. *Journal of Statistical Planning and Inference* **142**(8), 2241–2256.

474 T. Jiang and F. Yang (2013) Central limit theorems for classical likelihood ratio tests for high-dimensional normal distributions, *arXiv preprint arXiv:1306.0254*.

475 S. Péché (2009) Universality results for the largest eigenvalues of some sample covariance matrix ensembles. *Probability Theory and Related Fields* **143**(3–4), 481–516.

476 S. Chen, L. Zhang, and P. Zhong (2010) Tests for high-dimensional covariance matrices. *Journal of the American Statistical Association* **105**(490), 810–819.

477 M. S. Srivastava (2005) Some tests concerning the covariance matrix in high dimensional data. *J. Japan Statist. Soc* **35**(2), 251–272.

478 M. S. Srivastava (2007) Multivariate theory for analyzing high dimensional data. *J. Japan Statist. Soc* **37**(1), 53–86.

479 J. R. Schott (2005) Testing for complete independence in high dimensions. *Biometrika* **92**(4), 951–956.

480 J. Schott (2006) A high-dimensional test for the equality of the smallest eigenvalues of a covariance matrix. *Journal of Multivariate Analysis* **97**(4), 827–843.

481 J. R. Schott (2007) A test for the equality of covariance matrices when the dimension is large relative to the sample sizes. *Computational Statistics & Data Analysis* **51**(12), 6535–6542.

482 J. R. Schott (2010) Reduced-rank estimation of the difference between two covariance matrices. *Journal of Statistical Planning and Inference* **140**(4), 1038–1043.

483 Y. Fujikoshi, V. V. Ulyanov, and R. Shimizu (2011) *Multivariate statistics: High-dimensional and large-sample approximations* **760**. John Wiley & Sons, Ltd.

484 T. Cai and Z. Ma (2012) Optimal hypothesis testing for high dimensional covariance matrices, *arXiv preprint arXiv:1205.4219*.

485 C. Wang, L. Cao, and B. Miao (2013) Asymptotic power of likelihood ratio tests for high dimensional data, *arXiv preprint arXiv:1302.3302*.

486 C. Stein (1986) Lectures on the theory of estimation of many parameters. *Journal of Soviet Mathematics* **34**(1), 1373–1403.

487 E. Carter and M. Srivastava (1977) Monotonicity of the power functions of modified likelihood ratio criterion for the homogeneity of variances and of the sphericity test. *Journal of Multivariate Analysis* **7**(1), 229–233.

488 T. J. Fisher, X. Sun, and C. M. Gallagher (2010) A new test for sphericity of the covariance matrix for high dimensional data. *Journal of Multivariate Analysis* **101**(10), 2554–2570.

489 T. Cai, C. Zhang, and H. Zhou (2010) Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics* **38**(4), 2118–2144.

490 M. Birke and H. Dette (2005) A note on testing the covariance matrix for large dimension. *Statistics & probability letters* **74**(3), 281–289.

491 S. N. Roy and S. Roy (1957) *Some Aspects of Multivariate Analysis*, John Wiley & Sons, Inc.

492 H. Nagao (1973) On some test criteria for covariance matrix. *The Annals of Statistics*, 700–709.

493 J. W. Mauchly (1940) Significance test for sphericity of a normal n-variate distribution. *The Annals of Mathematical Statistics* **11**(2), 204–209.

494 L. J. Gleser (1966) A note on the sphericity test. *The Annals of Mathematical Statistics*, 464–467.

495 T. W. Anderson, T. W. Anderson, T. W. Anderson, and T. W. Anderson, (1958) *An introduction to multivariate statistical analysis*. 2nd ed. 1959, **66**(5).

496 B. Nagarsenker and K. Pillai (1973) The distribution of the sphericity test criterion. *Journal of Multivariate Analysis* **3**(2), 226–235.

497 S. John (1971) Some optimal multivariate tests. *Biometrika* **58**(1), 123–127.

498 S. John (1972) The distribution of a statistic used for testing sphericity of normal distributions. *Biometrika* **59**(1), 169–173.

499 F. Yang (2011) *Likelihood Ratio Tests for High-dimensional Normal Distributions*. Ph.D. dissertation. University of Minnesota.

500 Q. Wang and J. Yao (2013) On the sphericity test with large-dimensional observations, *arXiv preprint arXiv:1303.4035*.

501 S. Wilks (1935) On the independence of k sets of normally distributed statistical variables. *Econometrica, Journal of the Econometric Society*, 309–326.

502 M. S. Bartlett (1937) Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences* **160**(901), 268–282.

503 N. Sugiura and H. Nagao (1968) Unbiasedness of some test criteria for the equality of one or two covariance matrices. *The Annals of Mathematical Statistics* **39**(5), 1686–1692.

504 M. D. Perlman (1980) Unbiasedness of the likelihood ratio tests for equality of several covariance matrices and equality of several multivariate normal populations. *The Annals of Statistics* **8**(2), 247–263.

505 R. J. Schott (2001) Some tests for the equality of covariance matrices. *Journal of statistical planning and inference* **94**(1), 25–36.

506 L. Zhang (2013) *A Likelihood Ratio Test of Independence of Components for High-dimensional Normal Vectors*. Ph.D. thesis. University of Minnesota.

507 J. Gao, G. Pan, and M. Guo (2012) Independence Test for High Dimensional Random Vectors. Social Science Electronic Publishing, 2012. *Available at SSRN 2027295*.

508 Z. Bai and W. Zhou (2008) Large sample covariance matrices without independence structures in columns. *Statistica Sinica* **18**(2), 425.

509 Z. Bai and Y. Yin (1988) Convergence to the semicircle law. *The Annals of Probability* **16**(2), 863–875.

510 Z. Bai and J. Yao (2005) On the convergence of the spectral empirical process of wigner matrices. *Bernoulli* **11**(6), 1059–1092.

511 C. Wang, J. Yang, B. Miao, and L. Cao (2013) Identity tests for high dimensional data using RMT. *Journal of Multivariate Analysis*, 2013, **118**(5): 128–137.

512 A. Onatski, M. J. Moreira, and M. Hallin (2013) Asymptotic power of sphericity tests for high-dimensional data. *The Annals of Statistics* **41**(3), 1204–1231.

513 I. M. Johnstone and B. Nadler (2013) Roy's largest root test under rank-one alternatives, *arXiv preprint arXiv:1310.6581*.

514 R. R. Nadakuditi and J. W. Silverstein (2010) Fundamental limit of sample generalized eigenvalue based detection of signals in noise using relatively few signal-bearing and noise-only samples. *Selected Topics in Signal Processing, IEEE Journal of* **4**(3), 468–480.

515 R. V. Hogg, J. McKean, and A. T. Craig (2005) *Introduction to mathematical statistics*. Pearson Education.

516 H. L. Van Trees (1968) *Detection, Estimation, and Modulation Theory*, John Wiley & Sons, Inc.

517 O. Ryan and M. Debbah (2007) Free deconvolution for signal processing applications. *IEEE Trans. Information Theory* **1**, 1–15, Jan.

518 R. Vershynin (2011) Introduction to the non-asymptotic analysis of random matrices. *Arxiv preprint arXiv:1011.3027v5*, July.

519 R. Vershynin (2012) How close is the sample covariance matrix to the actual covariance matrix?. *Journal of Theoretical Probability* **25**(3), 655–686.

520 D. Li (2013) *Random Matrix Theory and Its Application in High-dimensional Statistics*. PhD thesis. University of Minnesota.

521 L. Haff (1980) Empirical Bayes estimation of the multivariate normal covariance matrix. *The Annals of Statistics* **8**(3), 586–597.

522 Q. Wang, J. W. Silverstein, and J. Yao (2013) A note on the clt of the lss for sample covariance matrix from a spiked population model, *arXiv preprint arXiv:1304.6164*.

523 B. Chen and G. Pan Clt for linear spectral statistics of normalized sample covariance matrices with larger dimension and small sample size, in *XXIX-th European Meeting of Statisticians, Budapest Contents*, p. 245.

524 J. Li, (2013) *Two sample inference for high dimensional data and nonparametric variable selection for census data*. PhD thesis, Iowa State University.

525 L. Huang and H. So (2013) Source enumeration via mdl criterion based on linear shrinkage estimation of noise subspace covariance matrix, *IEEE TRANSACTIONS ON SIGNAL PROCESSING* **61**, 4806–4821, October.

526 R. Couillet and E. Zio (2012) A subspace approach to fault diagnostics in large power systems, in *Communications Control and Signal Processing (ISCCSP), 2012 5th International Symposium on*, IEEE, 1–4.

527 EPRI, Epri intelligrid. http://smartgrid.epri.com/IntelliGrid.aspx,2016-12-11.

528 M. McGranaghan, D. Von Dollen, P. Myrda, and E. Gunther (2008) Utility experience with developing a smart grid roadmap, in *Power and Energy Society General Meeting-Conversion and Delivery of Electrical Energy in the 21st Century, 2008 IEEE*, IEEE, 1–5.

529 E. Commission (2006) European smartgrids technology platform. European Commission.

530 E. Commission (2005) Towards smart power networks.

531 E. Commission *et al.* (2006) European technology platform smart grids: vision and strategy for europe's electricity networks of the future. *Office for Official Publications of the European Communities*.

532 U. DOE (2003) Grid 2030: A national vision for electricity's second 100 years. *Department of Energy*.

533 NIST (2010) Nist framework and roadmap for smart grid interoperability standards, release 1.0. *National Institute of Standards and Technology*, 33.

534 X. Yu, C. Cecati, T. Dillon, and M. G. Simoes (2011) The new frontier of smart grids. *Industrial Electronics Magazine, IEEE* **5**(3), 49–63.

535 M. Liserre, T. Sauter, and J. Y. Hung (2010) Future energy systems: Integrating renewable energy sources into the smart power grid through industrial electronics. *Industrial Electronics Magazine, IEEE* **4**(1), 18–37.

536 W. A. Wulf (2000) Great achievements and grand challenges poised as we are between the twentieth and twenty-first centuries, it is the perfect moment to reflect on the accomplishments of engineers in the last century and ponder the challenges facing them in the next. *BRIDGE-WASHINGTON-* **30**(3/4), 5–10.

537 D. Von Dollen (2009) Report to nist on the smart grid interoperability standards roadmap. *Electric Power Research Institute (EPRI) and National Institute of Standards and Technology*.

538 H. Gharavi and R. Ghafurian (2011) Smart grid: The electric energy system of the future. *Proc. IEEE 99*(6), 917–921.

539 S. Massoud Amin and B. F. Wollenberg (2005) Toward a smart grid: power delivery for the 21st century. *Power and Energy Magazine, IEEE* **3**(5), 34–41.

540 A. Ipakchi and F. Albuyeh (2009) Grid of the future. *Power and Energy Magazine, IEEE* **7**(2), 52–62.

541 T. Garrity (2008) Getting smart. *Power and Energy Magazine, IEEE* **8**(2), 38–45.

542 H. Farhangi (2010) The path of the smart grid. *Power and Energy Magazine, IEEE* **8**(1), 18–28.

543 G. W. Arnold (2011) Challenges and opportunities in smart grid: A position article. *Proceedings of the IEEE* **99**(6), 922–927.

544 C.-C. Lin, C.-H. Yang, and J. Z. Shyua (2013) A comparison of innovation policy in the smart grid industry across the pacific: China and the USA. *Energy Policy*, 2013, **57**(7): 119–132.

545 S. Amin and A. M. Giacomoni (2012) Smart grid, safe grid. *Power and Energy Magazine, IEEE* **10**(1), 33–40.

546 A. Molderink, V. Bakker, M. G. Bosman, J. L. Hurink, and G. J. Smit (2010) Management and control of domestic smart grid technology. *Smart Grid, IEEE Transactions on* **1**(2), 109–119.

547 DOE (2006) Benefits of demand response in electricity markets and recommendations for achieving them ?a report to the united states congress pursuant to section 1252 of the energy policy act of 2005," tech. rep., Department of Energy.

548 A.-H. Mohsenian-Rad and A. Leon-Garcia (2010) Optimal residential load control with price prediction in real-time electricity pricing environments. *Smart Grid, IEEE Transactions on* **1**(2), 120–133.

549 A. R. Metke and R. L. Ekl (2010) Security technology for smart grid networks. *Smart Grid, IEEE Transactions on* **1**(1), 99–107.

550 C. Efthymiou and G. Kalogridis (2010) Smart grid privacy via anonymization of smart metering data, in *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference*, IEEE, 238–243.

551 T. Sauter and M. Lobashov (2011) End-to-end communication architecture for smart grids. *Industrial Electronics, IEEE Transactions on* **58**(4), 1218–1228.

552 V. C. Gungor, D. Sahin, T. Kocak, S. Ergut, C. Buccella, C. Cecati, and G. P. Hancke (2011) Smart grid technologies: communication technologies and standards. *Industrial informatics, IEEE transactions on* **7**(4), 529–539.

553 Leeds D J. The soft grid 2013–2020: Big data & utility analytics for smart grid. GTM Research, December 13, 2012, 7. "Market Trends–Electricity," US Energy Information Administration 8. Adam James, "How Capacity Markets Work," The Energy Collective, June 14, 2013, 9. Innovaro, TF2013-38, Fall 2013, 2012.

554 S. Rusitschka, K. Eger, and C. Gerdes (2010) Smart grid data cloud: A model for utilizing cloud computing in the smart grid domain, in *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on*, IEEE, 483–488.

555 J. G. Cupp and M. E. Beehler (2008) Implementing smart grid communications. *Burns and McDonnell Tech brief*.

556 M. Kezunovic (2011) Translational knowledge: from collecting data to making decisions in a smart grid. *Proceedings of the IEEE* **99**(6), 977–997.

557 P. P. Parikh, M. G. Kanabar, and T. S. Sidhu (2010) Opportunities and challenges of wireless communication technologies for smart grid applications, in *Power and Energy Society General Meeting, 2010 IEEE*, IEEE, 1–7.

558 Y. Zhu, G. Zhou, and Y. Zhu (2013) Present status and challenges of big data processing in smart grid (in chinese). *Power System Technology* **37**(4), 927–935.

559 IBM (2011) IBM Big Data Industry Energy & Utilities. Ibm Corporation.

560 D. Kligman (2012) Pg&es austin kicks off conference on dealing with smart grid data. http://www.pgecurrents.com/2012/08/14/pg-topic-is-dealing-with-data-that-comes-with-smart-grid/, 2016-12-11.

561 T. Groenfeldt (2012) Big data meets the smart electrical grid. http://www.forbes.com/sites/tomgroenfeldt/2012/05/09/big-data-meets-the-smart-electrical-grid/#631462ad1adc, 2016-12-11.

562 A. Pregelj, M. Begovic, and A. Rohatgi (2004) Quantitative techniques for analysis of large data sets in renewable distributed generation. *Power Systems, IEEE Transactions on* **19**(3), 1277–1285.

563 Vavilapalli V K, Murthy A C, Douglas C, et al. Apache hadoop yarn: Yet another resource negotiator. *Proceedings of the 4th annual Symposium on Cloud Computing*. ACM, 2013: 5.

564 Bigdata (2013) Big data challenges (in Chinese). *Electric Power IT* **11**(2), 5–10.

565 L. Nian, Y. Li, B. Li, and Z. Zhao (2013) Opportunity and challenge of big data for the power industry. *Electric Power Information Technology*, 2013, **11**(4): 1–4.

566 Z. Qu and S. Zhang (2012) The wams power data processing based on hadoop, in *2012 IACIT Hong Kong Conference*.

567 D. Wang, Y. Song, and Y. Zhu (2010) Information platform of smart grid based on cloud computing (in chinese). *Automation of Electric Power Systems* **34**(22), 7–12.

568 G. Heydt, C. Liu, A. Phadke, and V. Vittal (2001) Solution for the crisis in electric power supply. *Computer Applications in Power, IEEE* **14**(3), 22–30.

569 J. De La Ree, V. Centeno, J. S. Thorp, and A. G. Phadke (2010) Synchronized phasor measurement applications in power systems. *Smart Grid, IEEE Transactions on* **1**(1), 20–27.

570 R. F. Nuqui and A. G. Phadke (2005) Phasor measurement unit placement techniques for complete and incomplete observability. *Power Delivery, IEEE Transactions on* **20**(4), 2381–2388.

571 S. Chakrabarti, E. Kyriakides, and M. Albu (2009) Uncertainty in power system state variables obtained through synchronized measurements. *Instrumentation and Measurement, IEEE Transactions on* **58**(8), 2452–2458.

572 A. Bose (2010) Smart transmission grid applications and their supporting infrastructure. *Smart Grid, IEEE Transactions on* **1**(1), 11–19.

573 G. Mateos and G. B. Giannakis (2013) Load curve data cleansing and imputation via sparsity and low rank, *arXiv preprint arXiv:1301.7627*.

574 F. C. Schweppe, J. Wildes, and D. P. Rom (1970) Power system static-state estimation, parts i, ii, iii. *Power Apparatus and Systems, IEEE Transactions on* (1), 120–135.

575 Y. Huang, S. Werner, J. Huang, N. Kashyap, and V. Gupta (2012) State estimation in electric power grids. *IEEE Signal Processing Magazine*, 33–43.

576 F. F. Wu (1990) Power system state estimation: a survey. *International Journal of Electrical Power & Energy Systems* **12**(2), 80–87.

577 A. Monticelli (2000) Electric power system state estimation. *Proceedings of the IEEE* **88**(2), 262–282.

578 M. Zhou, V. A. Centeno, J. S. Thorp, and A. G. Phadke (2006) An alternative for including phasor measurements in state estimators. *Power Systems, IEEE Transactions on* **21**(4), 1930–1937.

579 V. Terzija, G. Valverde, D. Cai, P. Regulski, V. Madani, J. Fitch, S. Skok, M. M. Begovic, and A. Phadke (2011) Wide-area monitoring, protection, and control of future electric power networks. *Proceedings of the IEEE* **99**(1), 80–93.

580 A. Gomez-Exposito, A. Abur, A. de la Villa Jaen, and C. Gomez-Quiles (2011) A multilevel state estimation paradigm for smart grids. *Proceedings of the IEEE* **99**(6), 952–976.

581 M. Shahidehpour, H. Yamin, and Z. Li (2002) Market Operations in Electric Power Systems, New York, NY: IEEE. 2002.

582 F. C. Schweppe and E. J. Handschin (1974) Static state estimation in electric power systems. *Proceedings of the IEEE* **62**(7), 972–982.

583 A. Garcia, A. Monticelli, and P. Abreu (1979) Fast decoupled state estimation and bad data processing. *Power Apparatus and Systems, IEEE Transactions on* (5), 1645–1652.

584 J. Allemong, L. Radu, and A. Sasson (1982) A fast and reliable state estimation algorithm for aep's new control center. *Power Apparatus and Systems, IEEE Transactions on* (4), 933–944.

585 H. Zhu and G. B. Giannakis (2011) Estimating the state of ac power systems using semidefinite programming, in *North American Power Symposium (NAPS)*, 2011 IEEE, 1–7.

586 H. Zhu and G. B. Giannakis (2012) *Multi-area state estimation using distributed sdp for nonlinear power systems*, in *Smart Grid Communications (SmartGridComm), 2012 IEEE Third International Conference on*, IEEE, 623–628.

587 M. Brown Do Coutto Filho and J. S. de Souza (2009) Forecasting-aided state estimationart i: Panorama. *Power Systems, IEEE Transactions on* **24**(4), 1667–1677.

588 A. L. da Silva, M. Do Coutto Filho, and J. de Queiroz (1983) *State Forecasting in Electric Power Systems*, IEE Proceedings C (Generation, Transmission and Distribution), IET, vol. 130, pp. 237–244.

589 K. Moslehi and R. Kumar (2010) A reliability perspective of the smart grid. *Smart Grid, IEEE Transactions on* **1**(1), 57–64.

590 Y. Hu, A. Kuh, T. Yang, and A. Kavcic (2011) A belief propagation based power distribution system state estimator. *Computational Intelligence Magazine, IEEE* **6**(3), 36–46.

591 J. Zhu and A. Abur (2007) Bad data identification when using phasor measurements, in *Power Tech, 2007 IEEE Lausanne*, IEEE, 1676–1681.

592 O. Kosut, L. Jia, R. J. Thomas, and L. Tong (2011) Malicious data attacks on the smart grid. *Smart Grid, IEEE Transactions on* **2**(4), 645–658.

593 V. Kekatos and G. B. Giannakis (2012) Distributed robust power system state estimation. *IEEE Transactions on Power Systems*, 2013, **28**(2): 1617–1626.

594 W. Xu, M. Wang, and A. Tang (2011) Sparse recovery from nonlinear measurements with applications in bad data detection for power networks, *arXiv preprint arXiv:1112.6234*.

595 J. Chen and A. Abur (2005) Improved bad data processing via strategic placement of pmus, in *Power Engineering Society General Meeting, 2005 IEEE*, IEEE, 509–513.

596 J. Chen and A. Abur (2006) Placement of pmus to enable bad data detection in state estimation. *Power Systems, IEEE Transactions on* **21**(4), 1608–1615.

597 J. E. Tate and T. J. Overbye (2008) Line outage detection using phasor angle measurements. *Power Systems, IEEE Transactions on* **23**(4), 1644–1652.

598 T. Van Cutsem, M. Ribbens-Pavella, and L. Mili (1985) Bad data identification methods in power system state estimation-a comparative study. *Power Apparatus and Systems, IEEE Transactions on* (11), 3037–3049.

599 F. F. Wu and W.-H. Liu (1989) Detection of topology errors by state estimation [power systems]. *Power Systems, IEEE Transactions on* **4**(1), 176–183.

600 T. Van Cutsem, M. Ribbens-Pavella, and L. Mili (1984) Hypothesis testing identification: a new method for bad data analysis in power system state estimation. *Power Apparatus and Systems, IEEE Transactions on*, (11), 3239–3252.

601 Y. Liu, P. Ning, and M. K. Reiter (2011) False data injection attacks against state estimation in electric power grids. *ACM Transactions on Information and System Security (TISSEC)* **14**(1), 13.

602 A. J. Wood and B. F. Wollenberg (2012) *Power generation, operation, and control*. Wiley-Interscience.

603 A. Monticelli (1999) *State estimation in electric power systems: a generalized approach*, **507**. Springer.

604 L. Zhao and A. Abur (2005) Multi area state estimation using synchronized phasor measurements. *Power Systems, IEEE Transactions on* **20**(2), 611–617.

605 E. N. Asada, A. V. Garcia, and R. Romero (2005) Identifying multiple interacting bad data in power system state estimation, in *Power Engineering Society General Meeting, 2005 IEEE*, IEEE, 571–577.

606 S. Gastoni, G. Granelli, and M. Montagna (2003) Multiple bad data processing by genetic algorithms, in *Power Tech Conference Proceedings, 2003 IEEE Bologna*, vol. 1, pp. 6–pp, IEEE.

607 S. Kourouklis (1984) A large deviation result for the likelihood ratio statistic in exponential families. *The Annals of Statistics* **12**(4), 1510–1521.

608 D. Gorinevsky, S. Boyd, and S. Poll (2009) Estimation of faults in dc electrical power system, in *American Control Conference*, 2009, ACC'09, pp. 4334–4339, IEEE.

609 E. Handschin, F. Schweppe, J. Kohlas, and A. Fiechter (1975) Bad data analysis for power system state estimation. *Power Apparatus and Systems, IEEE Transactions on* **94**(2), 329–337.

610 B. Kavsin (1977) Widths of certain finite-dimensional sets and classes of smooth functions, Izv. AN SSSR, **41** (1977), 334–351. English transl. in Math. Izv, 1977, 11.

611 A. Y. Garnaev and E. D. Gluskin (1984) *The Widths of a Euclidean Ball*, Dokl. Akad. Nauk SSSR, **277**, pp. 1048–1052.

612 E. Candès and P. Randall (2006) Highly robust error correction by convex programming. Available at arxiv. org/abs. CS Patent 0,612,124.

613 J. Valenzuela, J. Wang, and N. Bissinger (2013) Real-time intrusion detection in power system operations. *IEEE Transactions on Power Systems* **28**, 1052–1062, May.

614 O. Kosut, L. Jia, R. J. Thomas, and L. Tong (2010) *Malicious Data Attacks on Smart Grid State Estimation: Attack Strategies and Countermeasures*, 2010 First IEEE International Conference, Smart Grid Communications (SmartGridComm), IEEE, pp. 220–225.

615 W. Wang and Z. Lu (2013) Cyber security in the Smart Grid: Survey and challenges. *Computer Networks*, 2013, **57**(5): 1344–1371.

616 H. Li, R. Mao, L. Lai, and R. C. Qiu (2010) *Compressed Meter Reading for Delay-Sensitive and Secure Load Report in Smart Grid*, 2010 First IEEE International Conference, Smart Grid Communications (SmartGridComm), IEEE, pp. 114–119.

617 F. Rahimi and A. Ipakchi (2010) Demand response as a market resource under the smart grid paradigm. *Smart Grid, IEEE Transactions on* **1**(1), 82–88.

618 A. J. Conejo, J. M. Morales, and L. Baringo (2010) Real-time demand response model. *Smart Grid, IEEE Transactions on* **1**(3), 236–242,

619 P. Samadi, A. Mohsenian-Rad, R. Schober, *et al.* (2010) *Optimal Real-Time Pricing Algorithm Based on Utility Maximization for Smart Grid*, 2010 First IEEE International Conference, Smart Grid Communications (SmartGridComm), IEEE, pp. 415–420.

620 A.-H. Mohsenian-Rad, V. W. Wong, J. Jatskevich, and R. Schober (2010) *Optimal and Autonomous Incentive-Based Energy Consumption Scheduling Algorithm for Smart Grid*, Innovative Smart Grid Technologies (ISGT), 2010, IEEE, pp. 1–6.

621 S. Deilami, A. S. Masoum, P. S. Moses, and M. A. Masoum (2011) Real-time coordination of plug-in electric vehicle charging in smart grids to minimize power losses and improve voltage profile. *Smart Grid, IEEE Transactions on* **2**(3), 456–467.

622 P. Wolfs (2010) *An Economic Assessment of second Use?lithium-ion Batteries for Grid Support*, 2010 20th Australasian, Universities Power Engineering Conference (AUPEC), IEEE, pp. 1–6.

623 S. Caron and G. Kesidis (2010) *Incentive-Based Energy Consumption Scheduling Algorithms for the Smart Grid*, 2010 First IEEE International Conference, Smart Grid Communications (SmartGridComm), IEEE, pp. 391–396.

624 S. Paudyal, C. A. Canizares, and K. Bhattacharya (2011) Optimal operation of distribution feeders in smart grids. *Industrial Electronics, IEEE Transactions on* **58**(10), 4495–4503.

625 A. Mohd, E. Ortjohann, A. Schmelter, *et al.* (2008) *Challenges in Integrating Distributed Energy Storage Systems into Future Smart Grid*, IEEE International Symposium, Industrial Electronics, 2008. ISIE 2008, IEEE, pp. 1627–1632.

626 H. Cramér (1970) *Random variables and probability distributions*. No. 36, Cambridge University Press.

627 D. Tse and P. Viswanath (2005) *Fundamentals of wireless communication*. Cambridge University Press.

628 S. Vishwanath, N. Jindal, and A. Goldsmith (2003) Duality, achievable rates, and sum-rate capacity of gaussian mimo broadcast channels. *Information Theory, IEEE Transactions on* **49**(10), 2658–2668.

629 H. Weingarten, Y. Steinberg, and S. Shamai (2006) The capacity region of the gaussian multiple-input multiple-output broadcast channel. *Information Theory, IEEE Transactions on* **52**(9), 3936–3964.

630 F. Penna and S. Stanczak (2012) *Decentralized Largest Eigenvalue Test for Multi-Sensor Signal Detection*, 2012 IEEE, Global Communications Conference (GLOBECOM), IEEE, pp. 3893–3898.

631 S. Stanczak, M. Goldenbaum, R. L. Cavalcante, and F. Penna (2012) *On In-Network Computation Via Wireless Multiple-Access Channels with Applications*, 2012 International Symposium, Wireless Communication Systems (ISWCS), IEEE, pp. 276–280.

632 F. Penna and S. Stanczak (2012) *Eigenvalue-Based Signal Detection in Cognitive Femtocell Networks Using a Decentralized Lanczos Algorithm*, 2012 IEEE International Symposium, Dynamic Spectrum Access Networks (DYSPAN), IEEE, pp. 283–283.

633 P. Zhang, R. Qiu, and N. Guo (2011) Demonstration of spectrum sensing with blindly learned features. *Communications Letters, IEEE* **15**(99), 548–550.

634 H. Khurana, M. Hadley, N. Lu, and D. A. Frincke (2010) Smart-grid security issues. *Security & Privacy, IEEE* **8**(1), 81–85.

635 A. Usman and S. H. Shami (2013) Evolution of communication technologies for smart grid applications. *Renewable and Sustainable Energy Reviews* **19**, 191–199.

636 N. Golmie, A. Scaglione, L. Lampe, and E. Yeh (2012) Guest editorial-smart grid communications. *Selected Areas in Communications, IEEE Journal on* **30**(6), 1025–1026.

637 F. Rusek, D. Persson, B. K. Lau, *et al.* (2013) Scaling up mimo: Opportunities and challenges with very large arrays. *Signal Processing Magazine, IEEE* **30**(1), 40–60.

638 H. Ngo, E. Larsson, and T. Marzetta (2011) Energy and spectral efficiency of very large multiuser mimo systems. *IEEE Transactions on Communications* **61**(4), 1436–1448.

639 G.-M. Pan, M.-H. Guo, and W. Zhou (2007) Asymptotic distributions of the signal-to-interference ratios of lmmse detection in multiuser communications. *The Annals of Applied Probability* **17**(1), 181–206.

640 Z. Bai and J. W. Silverstein (2007) On the signal-to-interference ratio of cdma systems in wireless communications. *The Annals of Applied Probability* **17**(1), 81–101.

641 V. C. Gungor, B. Lu, and G. P. Hancke (2010) Opportunities and challenges of wireless sensor networks in smart grid. *Industrial Electronics, IEEE Transactions on* **57**(10), 3557–3564.

642 F. Penna and S. Stanczak (2013) Decentralized eigenvalue algorithms for distributed signal detection in cognitive networks, *arXiv preprint arXiv:1303.7103*.

643 A. Giridhar and P. Kumar (2006) Toward a theory of in-network computation in wireless sensor networks. *Communications Magazine, IEEE* **44**(4), 98–107.
644 A. Giridhar and P. Kumar (2005) Computing and communicating functions over sensor networks. *Selected Areas in Communications, IEEE Journal on* **23**(4), 755–764.
645 L. Xiao and S. Boyd (2004) Fast linear iterations for distributed averaging. *Systems & Control Letters* **53**(1), 65–78.
646 C. D. Godsil, G. Royle, and C. Godsil (2001) *Algebraic graph theory* **207**. Springer New York.
647 S. Sardellitti, S. Barbarossa, and A. Swami (2012) Optimal topology control and power allocation for minimum energy consumption in consensus networks. *Signal Processing, IEEE Transactions on* **60**(1), 383–399.
648 S. Rai (2007) The spectrum of a random geometric graph is concentrated. *Journal of Theoretical Probability* **20**(2), 119–132.
649 C. Bordenave (2008) Eigenvalues of euclidean random matrices. *Random Structures & Algorithms* **33**(4), 515–532.
650 M. Mézard, G. Parisi, and A. Zee (1999) Spectra of euclidean random matrices. *Nuclear Physics B* **559**(3), 689–701.
651 X. Zeng, Distribution of eigenvalues of large Euclidean matrices generated from lp ellipsoid. *Statistics & Probability Letters*, 2014, 91: 181–191.
652 Zeng, X. A note on the large random inner-product kernel matrices. *Statistics & Probability Letters*, 2015, 99: 192–201.
653 R. Olfati-Saber, J. A. Fax, and R. M. Murray (2007) Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE* **95**(1), 215–233.
654 Z. Li, F. R. Yu, and M. Huang (2010) A distributed consensus-based cooperative spectrum-sensing scheme in cognitive radios. *Vehicular Technology, IEEE Transactions on* **59**(1), 383–393.
655 S. Kar and J. M. Moura (2009) Distributed consensus algorithms in sensor networks with imperfect communication: Link failures and channel noise. *IEEE Transactions on Signal Processing* **57**(1), 355–369.
656 I. D. Schizas, A. Ribeiro, and G. B. Giannakis (2008) Consensus in ad hoc WSNs with noisy links—Part I: Distributed estimation of deterministic signals. *IEEE Transactions on Signal Processing*, 2008, **56**(1): 350–364.
657 D. Li, S. Kar, J. Moura, H. V. Poor, and S. Cui (2014) Distributed kalman filtering over big data sets: Fundamental analysis through large deviations, *arXiv preprint arXiv:1402.0246*.
658 A. G. Dimakis, S. Kar, J. M. Moura, *et al.* (2010) Gossip algorithms for distributed signal processing. *Proceedings of the IEEE* **98**(11), 1847–1864.
659 M. Srivastava and C. Khatri (1979) *An Introduction to Multivariate Statistics*. North-Holland.
660 T. Siotani, Y. Fujikoshi, and T. Hayakawa (1985) *Modern multivariate statistical analysis, a graduate course and handbook*. American Sciences Press.
661 V. Kargin (2013) Lecture notes on free probability, *arXiv preprint arXiv:1305.2611*.

Index

Smart Grid using Big Data Analytics: A Random Matrix Theory Approach, First Edition.
Robert C. Qiu and Paul Antonik.