Springer Series in Statistics

For other titles published in this series, go to
http://www.springer.com/series/692

Zhidong Bai
Jack W. Silverstein

Spectral Analysis of Large Dimensional Random Matrices

Second Edition

Springer

Zhidong Bai
School of Mathematics and Statistics
KLAS MOE
Northeast Normal University
5268 Renmin Street
Changchun, Jilin 130024
China
baizd@nenu.edu.cn
&
Department of Statistics and Applied Probability
National University of Singapore
6 Science Drive 2
Singapore 117546
Singapore
stabaizd@nus.edu.sg

Jack W. Silverstein
Department of Mathematics
Box 8205
North Carolina State University
Raleigh, NC 27695-8205
jack@unity.ncsu.edu

ISSN 0172-7397
ISBN 978-1-4419-0660-1 e-ISBN 978-1-4419-0661-8
DOI 10.1007/978-1-4419-0661-8
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2009942423

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

This book is dedicated to:

Professor Calyampudi Radhakrishna Rao's 90th Birthday
Professor Ulf Grenander's 87th Birthday
Professor Yongquan Yin's 80th Birthday

and to

My wife, Xicun Dan, my sons
Li and Steve Gang, and grandsons
Yongji, and Yonglin

— Zhidong Bai

My children, Hila and Idan

— Jack W. Silverstein

Preface to the Second Edition

The ongoing developments being made in large dimensional data analysis continue to generate great interest in random matrix theory in both theoretical investigations and applications in many disciplines. This has doubtlessly contributed to the significant demand for this monograph, resulting in its first printing being sold out. The authors have received many requests to publish a second edition of the book.

Since the publication of the first edition in 2006, many new results have been reported in the literature. However, due to limitations in space, we cannot include all new achievements in the second edition. In accordance with the needs of statistics and signal processing, we have added a new chapter on the limiting behavior of eigenvectors of large dimensional sample covariance matrices. To illustrate the application of RMT to wireless communications and statistical finance, we have added a chapter on these areas. Certain new developments are commented on throughout the book. Some typos and errors found in the first edition have been corrected.

The authors would like to express their appreciation to Ms. Lü Hong for her help in the preparation of the second edition. They would also like to thank Professors Ying-Chang Liang, Zhaoben Fang, Baoxue Zhang, and Shurong Zheng, and Mr. Jiang Hu, for their valuable comments and suggestions. They also thank the copy editor, Mr. Hal Heinglein, for his careful reading, corrections, and helpful suggestions. The first author would like to acknowledge the support from grants NSFC 10871036, NUS R-155-000-079-112, and R-155-000-096-720.

Changchun, China, and Singapore — Zhidong Bai
Cary, North Carolina, USA — Jack W. Silverstein
March 2009

Preface to the First Edition

This monograph is an introductory book on the theory of random matrices (RMT). The theory dates back to the early development of quantum mechanics in the 1940s and 1950s. In an attempt to explain the complex organizational structure of heavy nuclei, E. Wigner, Professor of Mathematical Physics at Princeton University, argued that one should not compute energy levels from Schrödinger's equation. Instead, one should imagine the complex nuclei system as a black box described by $n \times n$ Hamiltonian matrices with elements drawn from a probability distribution with only mild constraints dictated by symmetry considerations. Under these assumptions and a mild condition imposed on the probability measure in the space of matrices, one finds the joint probability density of the n eigenvalues. Based on this consideration, Wigner established the well-known semicircular law. Since then, RMT has been developed into a big research area in mathematical physics and probability. Its rapid development can be seen from the following statistics from the Mathscinet database under keyword Random Matrix on 10 June 2005 (Table 0.1).

Table 0.1 Publication numbers on RMT in 10 year periods since 1955

1955–1964	1965–1974	1975–1984	1985–1994	1995–2004
23	138	249	635	1205

Modern developments in computer science and computing facilities motivate ever widening applications of RMT to many areas.

In statistics, classical limit theorems have been found to be seriously inadequate in aiding in the analysis of very high dimensional data.

In the biological sciences, a DNA sequence can be as long as several billion strands. In financial research, the number of different stocks can be as large as tens of thousands.

In wireless communications, the number of users can be several million.

All of these areas are challenging classical statistics. Based on these needs, the number of researchers on RMT is gradually increasing. The purpose of this monograph is to introduce the basic results and methodologies developed in RMT. We assume readers of this book are graduate students and beginning researchers who are interested in RMT. Thus, we are trying to provide the most advanced results with proofs using standard methods as detailed as we can.

After more than a half century, many different methodologies of RMT have been developed in the literature. Due to the limitation of our knowledge and length of the book, it is impossible to introduce all the procedures and results. What we shall introduce in this book are those results obtained either under moment restrictions using the moment convergence theorem or the Stieltjes transform.

In an attempt at complementing the material presented in this book, we have listed some recent publications on RMT that we have not introduced.

The authors would like to express their appreciation to Professors Chen Mufa, Lin Qun, and Shi Ningzhong, and Ms. Lü Hong for their encouragement and help in the preparation of the manuscript. They would also like to thank Professors Zhang Baoxue, Lee Sungchul, Zheng Shurong, Zhou Wang, and Hu Guorong for their valuable comments and suggestions.

Changchun, China — Zhidong Bai
Cary, North Carolina, USA — Jack W. Silverstein
June 2005

Contents

Chapter 1
Introduction

1.1 Large Dimensional Data Analysis

The aim of this book is to investigate the spectral properties of random matrices (RM) when their dimensions tend to infinity. All classical limiting theorems in statistics are under the assumption that the dimension of data is fixed. Then, it is natural to ask why the dimension needs to be considered large and whether there are any differences between the results for a fixed dimension and those for a large dimension.

In the past three or four decades, a significant and constant advancement in the world has been in the rapid development and wide application of computer science. Computing speed and storage capability have increased a thousand folds. This has enabled one to collect, store, and analyze data sets of very high dimension. These computational developments have had a strong impact on every branch of science. For example, Fisher's resampling theory had been silent for more than three decades due to the lack of efficient random number generators until Efron proposed his renowned bootstrap in the late 1970s; the minimum L_1 norm estimation had been ignored for centuries since it was proposed by Laplace until Huber revived it and further extended it to robust estimation in the early 1970s. It is difficult to imagine that these advanced areas in statistics would have received such deep development if there had been no assistance from the present-day computer.

Although modern computer technology helps us in so many respects, it also brings a new and urgent task to the statistician; that is, whether the classical limit theorems (i.e., those assuming a fixed dimension) are still valid for analyzing high dimensional data and how to remedy them if they are not.

Basically, there are two kinds of limiting results in multivariate analysis: those for a fixed dimension (classical limit theorems) and those for a large dimension (large dimensional limit theorems). The problem turns out to be which kind of result is closer to reality. As argued by Huber in [157], some statisticians might say that five samples for each parameter on average are

Z. Bai and J.W. Silverstein, *Spectral Analysis of Large Dimensional Random Matrices, Second Edition*, Springer Series in Statistics, DOI 10.1007/978-1-4419-0661-8_1,

enough to use asymptotic results. Now, suppose there are $p = 20$ parameters and we have a sample of size $n = 100$. We may consider the case as $p = 20$ being fixed and n tending to infinity, $p = 2\sqrt{n}$, or $p = 0.2n$. So, we have at least three different options from which to choose for an asymptotic setup. A natural question is then which setup is the best choice among the three. Huber strongly suggested studying the situation of an increasing dimension together with the sample size in linear regression analysis.

This situation occurs in many cases. In parameter estimation for a structured covariance matrix, simulation results show that parameter estimation becomes very poor when the number of parameters is more than four. Also, it is found in linear regression analysis that if the covariates are random (or have measurement errors) and the number of covariates is larger than six, the behavior of the estimates departs far away from the theoretic values unless the sample size is very large. In signal processing, when the number of signals is two or three and the number of sensors is more than 10, the traditional MUSIC (MUltiple SIgnal Classification) approach provides very poor estimation of the number of signals unless the sample size is larger than 1000. Paradoxically, if we use only half of the data set—namely, we use the data set collected by only five sensors—the signal number estimation is almost 100% correct if the sample size is larger than 200. Why would this paradox happen? Now, if the number of sensors (the dimension of data) is p, then one has to estimate p^2 parameters ($\frac{1}{2}p(p+1)$ real parts and $\frac{1}{2}p(p-1)$ imaginary parts of the covariance matrix). Therefore, when p increases, the number of parameters to be estimated increases proportional to p^2 while the number $(2np)$ of observations increases proportional to p. This is the underlying reason for this paradox. This suggests that one has to revise the traditional MUSIC method if the sensor number is large.

An interesting problem was discussed by Bai and Saranadasa [27], who theoretically proved that when testing the difference of means of two high dimensional populations, Dempster's [91] nonexact test is more powerful than Hotelling's T^2 test even when the T^2 statistic is well defined.

It is well known that statistical efficiency will be significantly reduced when the dimension of data or number of parameters becomes large. Thus, several techniques for dimension reduction have been developed in multivariate statistical analysis. As an example, let us consider a problem in principal component analysis. If the data dimension is 10, one may select three principal components so that more than 80% of the information is reserved in the principal components. However, if the data dimension is 1000 and 300 principal components are selected, one would still have to face a high dimensional problem. If one only chooses three principal components, he would have lost 90% or even more of the information carried in the original data set. Now, let us consider another example.

Example 1.1. Let X_{ij} be iid standard normal variables. Write

$$S_n = \left(\frac{1}{n}\sum_{k=1}^{n} X_{ik}X_{jk}\right)_{i,j=1}^{p},$$

which can be considered as a sample covariance matrix with n samples of a p-dimensional mean-zero random vector with population matrix I. An important statistic in multivariate analysis is

$$T_n = \log(\det S_n) = \sum_{j=1}^{p} \log(\lambda_{n,j}),$$

where $\lambda_{n,j}$, $j = 1, \cdots, p$, are the eigenvalues of S_n. When p is fixed, $\lambda_{n,j} \to 1$ almost surely as $n \to \infty$ and thus $T_n \xrightarrow{\text{a.s.}} 0$.

Further, by taking a Taylor expansion on $\log(1 + x)$, one can show that

$$\sqrt{n/p}\, T_n \xrightarrow{\mathscr{D}} N(0, 2),$$

for any fixed p. This suggests the possibility that T_n is asymptotically normal, provided that $p = O(n)$. However, this is not the case. Let us see what happens when $p/n \to y \in (0, 1)$ as $n \to \infty$. Using results on the limiting spectral distribution of $\{S_n\}$ (see Chapter 3), we will show that with probability 1

$$\frac{1}{p}T_n \to \int_{a(y)}^{b(y)} \frac{\log x}{2\pi xy}\sqrt{(b(y)-x)(x-a(y))}dx = \frac{y-1}{y}\log(1-y)-1 \equiv d(y) < 0 \tag{1.1.1}$$

where $a(y) = (1 - \sqrt{y})^2$, $b(y) = (1 + \sqrt{y})^2$. This shows that almost surely

$$\sqrt{n/p}\, T_n \sim d(y)\sqrt{np} \to -\infty.$$

Thus, any test that assumes asymptotic normality of T_n will result in a serious error.

These examples show that the classical limit theorems are no longer suitable for dealing with high dimensional data analysis. Statisticians must seek out special limiting theorems to deal with large dimensional statistical problems. Thus, the theory of random matrices (RMT) might be one possible method for dealing with large dimensional data analysis and hence has received more attention among statisticians in recent years. For the same reason, the importance of RMT has found applications in many research areas, such as signal processing, network security, image processing, genetic statistics, stock market analysis, and other finance or economic problems.

1.2 Random Matrix Theory

RMT traces back to the development of quantum mechanics (QM) in the 1940s and early 1950s. In QM, the energy levels of a system are described by eigenvalues of a Hermitian operator **A** on a Hilbert space, called the Hamiltonian. To avoid working with an infinite dimensional operator, it is common to approximate the system by discretization, amounting to a truncation, keeping only the part of the Hilbert space that is important to the problem under consideration. Hence, the limiting behavior of large dimensional random matrices has attracted special interest among those working in QM, and many laws were discovered during that time. For a more detailed review on applications of RMT in QM and other related areas, the reader is referred to the book *Random Matrices* by Mehta [212].

Since the late 1950s, research on the limiting spectral analysis of large dimensional random matrices has attracted considerable interest among mathematicians, probabilists, and statisticians. One pioneering work is the semicircular law for a Gaussian (or Wigner) matrix (see Chapter 2 for the definition), due to Wigner [296, 295]. He proved that the expected spectral distribution of a large dimensional Wigner matrix tends to the so-called semicircular law. This work was generalized by Arnold [8, 7] and Grenander [136] in various aspects. Bai and Yin [37] proved that the spectral distribution of a sample covariance matrix (suitably normalized) tends to the semicircular law when the dimension is relatively smaller than the sample size. Following the work of Marčenko and Pastur [201] and Pastur [230, 229], the asymptotic theory of spectral analysis of large dimensional sample covariance matrices was developed by many researchers, including Bai, Yin, and Krishnaiah [41], Grenander and Silverstein [137], Jonsson [169], Wachter [291, 290], Yin [300], and Yin and Krishnaiah [304]. Also, Yin, Bai, and Krishnaiah [301, 302], Silverstein [260], Wachter [290], Yin [300], and Yin and Krishnaiah [304] investigated the limiting spectral distribution of the multivariate F-matrix, or more generally of products of random matrices. In the early 1980s, major contributions on the existence of the limiting spectral distribution (LSD) and their explicit forms for certain classes of random matrices were made. In recent years, research on RMT has turned toward second-order limiting theorems, such as the central limit theorem for linear spectral statistics, the limiting distributions of spectral spacings, and extreme eigenvalues.

1.2.1 Spectral Analysis of Large Dimensional Random Matrices

Suppose **A** is an $m\times m$ matrix with eigenvalues λ_j, $j = 1, 2, \cdots, m$. If all these eigenvalues are real (e.g., if **A** is Hermitian), we can define a one-dimensional distribution function

$$F^{\mathbf{A}}(x) = \frac{1}{m}\#\{j \le m : \lambda_j \le x\} \tag{1.2.1}$$

called the empirical spectral distribution (ESD) of the matrix $\mathbf{A}$. Here $\#E$ denotes the cardinality of the set E. If the eigenvalues λ_j's are not all real, we can define a two-dimensional empirical spectral distribution of the matrix $\mathbf{A}$:

$$F^{\mathbf{A}}(x, y) = \frac{1}{m}\#\{j \le m : \Re(\lambda_j) \le x, \Im(\lambda_j) \le y\}. \tag{1.2.2}$$

One of the main problems in RMT is to investigate the convergence of the sequence of empirical spectral distributions $\{F^{\mathbf{A}_n}\}$ for a given sequence of random matrices $\{\mathbf{A}_n\}$. The limit distribution F (possibly defective; that is, total mass is less than 1 when some eigenvalues tend to $\pm\infty$), which is usually nonrandom, is called the *limiting spectral distribution* (LSD) of the sequence $\{\mathbf{A}_n\}$.

We are especially interested in sequences of random matrices with dimension (number of columns) tending to infinity, which refers to *the theory of large dimensional random matrices.*

The importance of ESD is due to the fact that many important statistics in multivariate analysis can be expressed as functionals of the ESD of some RM. We now give a few examples.

Example 1.2. Let $\mathbf{A}$ be an $n \times n$ positive definite matrix. Then

$$\det(\mathbf{A}) = \prod_{j=1}^{n} \lambda_j = \exp\left(n \int_0^\infty \log x F^{\mathbf{A}}(dx)\right).$$

Example 1.3. Let the covariance matrix of a population have the form $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_q + \sigma^2\mathbf{I}$, where the dimension of $\boldsymbol{\Sigma}$ is p and the rank of $\boldsymbol{\Sigma}_q$ is $q(< p)$. Suppose $\mathbf{S}$ is the sample covariance matrix based on n iid samples drawn from the population. Denote the eigenvalues of $\mathbf{S}$ by $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_p$. Then the test statistic for the hypothesis H_0 : $\mathrm{rank}(\boldsymbol{\Sigma}_q) = q$ against H_1 : $\mathrm{rank}(\boldsymbol{\Sigma}_q) > q$ is given by

$$T = \frac{1}{p-q}\sum_{j=q+1}^{p} \sigma_j^2 - \left(\frac{1}{p-q}\sum_{j=q+1}^{p} \sigma_j\right)^2$$

$$= \frac{p}{p-q}\int_0^{\sigma_q} x^2 F^{\mathbf{S}}(dx) - \left(\frac{p}{p-q}\int_0^{\sigma_q} x F^{\mathbf{S}}(dx)\right)^2.$$

1.2.2 Limits of Extreme Eigenvalues

In applications of the asymptotic theorems of spectral analysis of large dimensional random matrices, two important problems arise after the LSD is found. The first is the bound on extreme eigenvalues; the second is the convergence rate of the ESD with respect to sample size. For the first problem, the literature is extensive. The first success was due to Geman [118], who proved that the largest eigenvalue of a sample covariance matrix converges almost surely to a limit under a growth condition on all the moments of the underlying distribution. Yin, Bai, and Krishnaiah [301] proved the same result under the existence of the fourth moment, and Bai, Silverstein, and Yin [33] proved that the existence of the fourth moment is also necessary for the existence of the limit. Bai and Yin [38] found the necessary and sufficient conditions for almost sure convergence of the largest eigenvalue of a Wigner matrix. By the symmetry between the largest and smallest eigenvalues of a Wigner matrix, the necessary and sufficient conditions for almost sure convergence of the smallest eigenvalue of a Wigner matrix were also found.

Compared to almost sure convergence of the largest eigenvalue of a sample covariance matrix, a relatively harder problem is to find the limit of the smallest eigenvalue of a large dimensional sample covariance matrix. The first attempt was made in Yin, Bai, and Krishnaiah [302], in which it was proved that the almost sure limit of the smallest eigenvalue of a Wishart matrix has a positive lower bound when the ratio of the dimension to the degrees of freedom is less than 1/2. Silverstein [262] modified the work to allow a ratio less than 1. Silverstein [263] further proved that, with probability 1, the smallest eigenvalue of a Wishart matrix tends to the lower bound of the LSD when the ratio of the dimension to the degrees of freedom is less than 1. However, Silverstein's approach strongly relies on the normality assumption on the underlying distribution and thus cannot be extended to the general case. The most current contribution was made in Bai and Yin [36], in which it is proved that, under the existence of the fourth moment of the underlying distribution, the smallest eigenvalue (when $p \le n$) or the $p-n+1$st smallest eigenvalue (when $p > n$) tends to $a(y) = \sigma^2(1-\sqrt{y})^2$, where $y = \lim(p/n) \in (0, \infty)$. Compared to the case of the largest eigenvalues of a sample covariance matrix, the existence of the fourth moment seems to be necessary also for the problem of the smallest eigenvalue. However, this problem has not yet been solved.

1.2.3 Convergence Rate of the ESD

The second problem, the convergence rate of the spectral distributions of large dimensional random matrices, is of practical interest. Indeed, when the LSD is used in estimating functionals of eigenvalues of a random matrix, it is important to understand the reliability of performing the substitution. This problem had been open for decades. In finding the limits of both the LSD and the extreme eigenvalues of symmetric random matrices, a very useful and powerful method is the moment method, which does not give any information about the rate of the convergence of the ESD to the LSD. The first success was made in Bai [16, 17], in which a Berry-Esseen type inequality of the difference of two distributions was established in terms of their Stieltjes transforms. Applying this inequality, a convergence rate for the expected ESD of a large Wigner matrix was proved to be $O(n^{-1/4})$ and that for the sample covariance matrix was shown to be $O(n^{-1/4})$ if the ratio of the dimension to the degrees of freedom is far from 1 and $O(n^{-5/48})$ if the ratio is close to 1. Some further developments can be found in Bai et al. [23, 24, 25], Bai et al. [26], Götze et al. [132], and Götze and Tikhomirov [133, 134].

1.2.4 Circular Law

The most perplexing problem is the so-called circular law, which conjectures that the spectral distribution of a nonsymmetric random matrix, after suitable normalization, tends to the uniform distribution over the unit disk in the complex plane. The difficulty exists in that two of the most important tools used for symmetric matrices do not apply for nonsymmetric matrices. Furthermore, certain truncation and centralization techniques cannot be used. The first known result was given in Mehta [212] (1967 edition) and in an unpublished paper of Silverstein (1984) that was reported in Hwang [159]. They considered the case where the entries of the matrix are iid standard complex normal. Their method uses the explicit expression of the joint density of the complex eigenvalues of the random matrix that was found by Ginibre [120]. The first attempt to prove this conjecture under some general conditions was made in Girko [123, 124]. However, his proofs contain serious mathematical gaps and have been considered questionable in the literature. Recently, Edelman [98] found the conditional joint distribution of complex eigenvalues of a random matrix whose entries are real normal $N(0, 1)$ when the number of its real eigenvalues is given and proved that the expected spectral distribution of the real Gaussian matrix tends to the circular law. Under the existence of the $4 + \varepsilon$ moment and the existence of a density, Bai [14] proved the strong version of the circular law. Recent work has eliminated the density requirement and weakened the moment condition. Further details are given in Chapter 11. Some consequent achievements can be found in Pan and Zhou [227] and Tao and Vu [273].

1.2.5 CLT of Linear Spectral Statistics

As mentioned above, functionals of the ESD of RMs are important in multivariate inference. Indeed, a parameter θ of the population can sometimes be expressed as

$$\theta = \int f(x)dF(x).$$

To make statistical inference on θ, one may use the integral

$$\hat{\theta} = \int f(x)dF_n(x),$$

which we call *linear spectral statistics* (LSS), as an estimator of θ, where $F_n(x)$ is the ESD of the RM computed from the data set. Further, one may want to know the limiting distribution of $\hat{\theta}$ through suitable normalization. In Bai and Silverstein [30], the normalization was found to be n by showing the limiting distribution of the linear functional

$$X_n(f) = n \int f(t)d(F_n(t) - F(t))$$

to be Gaussian under certain assumptions.

The first work in this direction was done by Jonsson [169], in which $f(t) = t^r$ and F_n is the ESD of a normalized standard Wishart matrix. Further work was done by Johansson [165], Bai and Silverstein [30], Bai and Yao [35], Sinai and Soshnikov [269], Anderson and Zeitouni [2], and Chatterjee [77], among others.

It would seem natural to pursue the properties of linear functionals by way of proving results on the process $G_n(t) = \alpha_n(F_n(t) - F(t))$ when viewed as a random element in $D[0,\infty)$, the metric space of functions with discontinuities of the first kind, along with the Skorohod metric. Unfortunately, this is impossible. The work done in Bai and Silverstein [30] shows that $G_n(t)$ cannot converge weakly to any nontrivial process for any choice of α_n. This fact appears to occur in other random matrix ensembles. When F_n is the empirical distribution of the angles of eigenvalues of an $n \times n$ Haar matrix, Diaconis and Evans [94] proved that all finite dimensional distributions of $G_n(t)$ converge in distribution to independent Gaussian variables when $\alpha_n = n/\sqrt{\log n}$. This shows that with $\alpha_n = n/\sqrt{\log n}$, the process G_n cannot be tight in $D[0,\infty)$.

The result of Bai and Silverstein [30] has been applied in several areas, especially in wireless communications, where sample covariance matrices are used to model transmission between groups of antennas. See, for example, Tulino and Verdu [283] and Kamath and Hughes [170].

1.2.6 Limiting Distributions of Extreme Eigenvalues and Spacings

The first work on the limiting distributions of extreme eigenvalues was done by Tracy and Widom [278], who found the expression for the largest eigenvalue of a Gaussian matrix when suitably normalized. Further, Johnstone [168] found the limiting distribution of the largest eigenvalue of the large Wishart matrix. In El Karoui [101], the Tracy-Widom law of the largest eigenvalue is established for the complex Wishart matrix when the population covariance matrix differs from the identity.

When the majority of the population eigenvalues are 1 and some are larger than 1, Johnstone proposed the *spiked eigenvalues model* in [168]. Then, Baik et al. [43] and Baik and Silverstein [44] investigated the strong limit of spiked eigenvalues. Bai and Yao [34] investigated the CLT of spiked eigenvalues. A special case of the CLT when the underlying distribution is complex Gaussian was considered in Baik et al. [43], and the real Gaussian case was considered in Paul [231].

The work on spectrum spacing has a long history that dates back to Mehta [213]. Most of the work in these two directions assumes the Gaussian (or generalized) distributions.

1.3 Methodologies

The eigenvalues of a matrix can be regarded as continuous functions of entries of the matrix. But these functions have no closed form when the dimension of the matrix is larger than 4. So special methods are needed to understand them. There are three important methods employed in this area: the moment method, Stieltjes transform, and orthogonal polynomial decomposition of the exact density of eigenvalues. Of course, the third method needs the assumption of the existence and special forms of the densities of the underlying distributions in the RM.

1.3.1 Moment Method

In the following, $\{F_n\}$ will denote a sequence of distribution functions, and the k-th moment of the distribution F_n is denoted by

$$\beta_{n,k} = \beta_k(F_n) := \int x^k dF_n(x). \tag{1.3.1}$$

The moment method is based on the moment convergence theorem (MCT); see Lemmas B.1, B.2, and B.3.

Let $\mathbf{A}$ be an $n \times n$ Hermitian matrix, and denote its eigenvalues by $\lambda_1 \leq \cdots \leq \lambda_n$. The ESD, $F^{\mathbf{A}}$, of $\mathbf{A}$ is defined as in (1.2.1) with m replaced by n. Then, the k-th moment of $F^{\mathbf{A}}$ can be written as

$$\beta_{n,k}(\mathbf{A}) = \int_{-\infty}^{\infty} x^k F^{\mathbf{A}}(dx) = \frac{1}{n}\mathrm{tr}(\mathbf{A}^k). \tag{1.3.2}$$

This expression plays a fundamental role in RMT. By MCT, the problem of showing that the ESD of a sequence of random matrices $\{\mathbf{A}_n\}$ (strongly or weakly or in another sense) tends to a limit reduces to showing that, for each fixed k, the sequence $\{\frac{1}{n}\mathrm{tr}(\mathbf{A}^k)\}$ tends to a limit β_k in the corresponding sense and then verifying the Carleman condition (B.1.4),

$$\sum_{k=1}^{\infty} \beta_{2k}^{-1/2k} = \infty.$$

Note that in most cases the LSD has finite support, and hence the characteristic function of the LSD is analytic and the necessary condition for the MCT holds automatically. Most results in finding the LSD or proving the existence of the LSD were obtained by estimating the mean, variance, or higher moments of $\frac{1}{n}\mathrm{tr}(\mathbf{A}^k)$.

1.3.2 Stieltjes Transform

The definition and simple properties of the Stieltjes transform can be found in Appendix B, Section B.2. Here, we just illustrate how it can be used in RMT. Let $\mathbf{A}$ be an $n \times n$ Hermitian matrix and F_n be its ESD. Then, the Stieltjes transform of F_n is given by

$$s_n(z) = \int \frac{1}{x - z} dF_n(x) = \frac{1}{n}\mathrm{tr}(\mathbf{A} - z\mathbf{I})^{-1}.$$

Using the inverse matrix formula (see Theorem A.4), we get

$$s_n(z) = \frac{1}{n} \sum_{k=1}^{n} \frac{1}{a_{kk} - z - \boldsymbol{\alpha}_k^*(\mathbf{A}_k - z\mathbf{I})^{-1}\boldsymbol{\alpha}_k}$$

where $\mathbf{A}_k$ is the $(n-1) \times (n-1)$ matrix obtained from $\mathbf{A}$ with the k-th row and column removed and $\boldsymbol{\alpha}_k$ is the k-th column vector of $\mathbf{A}$ with the k-th element removed.

If the denominator $a_{kk} - z - \boldsymbol{\alpha}_k^*(\mathbf{A}_k - z\mathbf{I})^{-1}\boldsymbol{\alpha}_k$ can be proven to be equal to $g(z, s_n(z)) + o(1)$ for some function g, then the LSD F exists and its Stieltjes transform of F is the solution to the equation

$$s = 1/g(z, s).$$

Its applications will be discussed in more detail later.

1.3.3 Orthogonal Polynomial Decomposition

Assume that the matrix $\mathbf{A}$ has a density $p_n(\mathbf{A}) = H(\lambda_1, \cdots, \lambda_n)$. It is known that the joint density function of the eigenvalues will be of the form

$$p_n(\lambda_1, \cdots, \lambda_n) = cJ(\lambda_1, \cdots, \lambda_n)H(\lambda_1, \cdots, \lambda_n),$$

where J comes from the integral of the Jacobian of the transform from the matrix space to its eigenvalue-eigenvector space. Generally, it is assumed that H has the form $H(\lambda_1, \cdots, \lambda_n) = \prod_{k=1}^n g(\lambda_k)$ and J has the form $\prod_{i<j}(\lambda_i - \lambda_j)^\beta \prod_{k=1}^n h_n(\lambda_k)$. 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 $\beta = 1$ and $h_n(x) = x^{n-p}$ for a real Wishart matrix with $n \geq p$.

Examples considered in the literature are the following

(1) Real Gaussian matrix (symmetric; i.e., $\mathbf{A}' = \mathbf{A}$):

$$p_n(\mathbf{A}) = c \exp\left(-\frac{1}{4\sigma^2}\operatorname{tr}(\mathbf{A}^2)\right).$$

In this case, the diagonal entries of $\mathbf{A}$ are iid real $N(0, 2\sigma^2)$ and entries above diagonal are iid real $N(0, \sigma^2)$.

(2) Complex Gaussian matrix (Hermitian; i.e., $\mathbf{A}^* = \mathbf{A}$):

$$p_n(\mathbf{A}) = c \exp\left(-\frac{1}{2\sigma^2}\operatorname{tr}(\mathbf{A}^2)\right).$$

In this case, the diagonal entries of $\mathbf{A}$ are iid real $N(0, \sigma^2)$ and entries above diagonal are iid complex $N(0, \sigma^2)$ (whose real and imaginary parts are iid $N(0, \sigma^2/2)$).

(3) Real Wishart matrix of order $p \times n$:

$$p_n(\mathbf{A}) = c \exp\left(-\frac{1}{2\sigma^2}\operatorname{tr}(\mathbf{A}'\mathbf{A})\right).$$

In this case, the entries of $\mathbf{A}$ are iid real $N(0, \sigma^2)$.

(4) Complex Wishart matrix of order $p \times n$:

$$p_n(\mathbf{A}) = c \exp\left(-\frac{1}{\sigma^2}\mathrm{tr}(\mathbf{A}^*\mathbf{A})\right).$$

In this case, the entries of $\mathbf{A}$ are iid complex $N(0, \sigma^2)$.

For generalized densities, there are the following.

(1) Symmetric matrix:

$$p_n(\mathbf{A}) = c \exp(-\mathrm{tr}V(\mathbf{A})).$$

(2) Hermitian matrix:

$$p_n(\mathbf{A}) = c \exp(-\mathrm{tr}V(\mathbf{A})).$$

In the two cases above, V is assumed to be a polynomial of even degree with a positive leading coefficient.

(3) Real covariance matrix of dimension p and degrees of freedom n:

$$p_n(\mathbf{A}) = c \exp(-\mathrm{tr}V(\mathbf{A}'\mathbf{A})).$$

(4) Complex covariance matrix of dimension p and degrees of freedom n:

$$p_n(\mathbf{A}) = c \exp(-\mathrm{tr}V(\mathbf{A}^*\mathbf{A})).$$

In the two cases above, V is assumed to be a polynomial with a positive leading coefficient.

Note that the factor $\prod_{i<j}(\lambda_i - \lambda_j)$ is the determinant of the Vandermonde matrix generated by $\lambda_1, \cdots, \lambda_n$. Therefore, we may rewrite the density of the eigenvalues of the matrices as

$$\begin{aligned}
&p_n(\lambda_1, \cdots, \lambda_n) \\
&= c \prod_{k=1}^{n} h_n(\lambda_k) g(\lambda_k) \det \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_2 & \cdots & \lambda_n \\ \vdots & \vdots & \cdots & \vdots \\ \lambda_1^{n-1} & \lambda_2^{n-1} & \cdots & \lambda_n^{n-1} \end{pmatrix}^{\beta} \\
&= c \prod_{k=1}^{n} h_n(\lambda_k) g(\lambda_k) \det \begin{pmatrix} 1 & 1 & \cdots & 1 \\ m_1(\lambda_1) & m_1(\lambda_2) & \cdots & m_1(\lambda_n) \\ \vdots & \vdots & \cdots & \vdots \\ m_{n-1}(\lambda_1) & m_{n-1}(\lambda_2) & \cdots & m_{n-1}(\lambda_n) \end{pmatrix}^{\beta},
\end{aligned}$$

where m_k is any polynomial of degree k and having leading coefficient 1. For ease of finding the marginal densities of several eigenvalues, one may choose the m functions as orthogonal polynomials with respective $[g(x)h_n(x)]^{2/\beta}$. Then, through mathematical analysis, one can draw various conclusions from the expression above.

Note that the moment method and Stieltjes transform method can be done under moment assumptions. This book will primarily concentrate on results without assuming density conditions. Readers who are interested in the method of orthogonal polynomials are referred to Deift [88].

1.3.4 Free Probability

Free probability is a mathematical theory that studies noncommutative random variables. The "freeness" property is the analogue of the classical notion of independence, and it is connected with free products. This theory was initiated by Dan Voiculescu around 1986 in order to attack the free group factors isomorphism problem, an important unsolved problem in the theory of operator algebras. Typically the random variables lies in a unital algebra A such as a C^* algebra or a von Neumann algebra. The algebra comes equipped with a noncommutative expectation, a linear functional $\varphi : A \to C$ such that $\varphi(1) = 1$. Unital subalgebras $A_1, \cdots, A_n$ are then said to be free if the expectation of the product $a_1 \cdots a_n$ is zero whenever each a_j has zero expectation, lies in an A_k, and no adjacent a_j's come from the same subalgebra A_k. Random variables are free if they generate free unital subalgebras.

An interesting aspect and active research direction of free probability lies in its applications to RMT. The functional φ stands for the normalized expected trace of a random matrix. For any $n \times n$ Hermitian random matrix $\mathbf{A}_n$ and a given integer k, $\varphi(\mathbf{A}_n^k) = \frac{1}{n}\mathrm{tr}(\mathrm{E}\mathbf{A}_n^k)$. If $\lim_n \varphi(\mathbf{A}_n^k) = \alpha_k$, for all k, then instead of referring to the collection of numbers α_k, it is better to use some random variable A (if it exists) to characterize the α_k's as moments of A. By setting $\varphi(A^k) = \alpha_k$, one may say that $\mathbf{A}_n \to A$ in distribution. A general definition is given as follows.

Definition 1.4. Consider $n \times n$ random matrices $A_n^{(1)}, \cdots, A_n^{(m)}$ and variables $A_1, \cdots, A_m$. We say that

$$(A_n^{(1)}, \cdots, A_n^{(m)}) \to (A_1, \cdots, A_m) \quad \text{in distribution}$$

if

$$\lim_{n\to\infty} \varphi(A_n^{(i_1)} \cdots A_n^{(i_k)}) = \varphi(A_{i_1} \cdots A_{i_k})$$

for all choices of k, $1 \le i_1, \cdots, i_k \le m$.

When $m = 1$, the definition of convergence in distribution is to say that if the normalized expected trace of $\mathbf{A}_n^k$ tends to the k-th moment of A, then we define $\mathbf{A}_n$ tending to A. For example, let $\mathbf{A}_n$ be the normalized Wigner matrix (see Chapter 2). Then A is the semicircular law. Now, suppose we have two independent sequences of normalized Wigner matrices, $\{\mathbf{A}_n\}$ and $\{\mathbf{B}_n\}$. How do we characterize their limits? If individually, then $\mathbf{A}_n \to s_a$ and $\mathbf{B}_n \to s_b$, and both s_a and s_b are semicircular laws. The problem is how to consider the joint limit of the sequences of pairs $(\mathbf{A}_n, \mathbf{B}_n)$. Or equivalently, what is the relationship of s_a and s_b? According to free probability, we have the following definition.

Definition 1.5. The matrices $\mathbf{A}_1, \cdots, \mathbf{A}_m$ are called free if

$$\varphi([p_1(\mathbf{A}_{i_1}) \cdots p_k(\mathbf{A}_{i_k})]) = 0$$

whenever

- $p_1, \cdots, p_k$ are polynomials in one variable,
- $i_1 \neq i_2 \neq i_3 \neq \cdots \neq i_k$ (only neighboring elements are required to be distinct),
- $\varphi(p_j(\mathbf{A}_{i_j})) = 0$ for all $j = 1, \cdots, k$.

Note that the definition of freeness can be considered as a way of organizing the information about all joint moments of free variables in a systematic and conceptual way. Indeed, the definition above allows one to calculate mixed moments of free variables in terms of moments of the single variables. For example, if a, b are free, then the definition of freeness requires that $\varphi[(a - \varphi(a)1)(b - \varphi(b)1)] = 0$, which implies that $\varphi(ab) = \varphi(a)\varphi(b)$. In the same way, $\varphi[(a - \varphi(a)1)(b - \varphi(b)1)(a - \varphi(a)1)(b - \varphi(b)1)] = 0$ leads finally to $\varphi(abab) = \varphi(aa)\varphi(b)\varphi(b) + \varphi(a)\varphi(a)\varphi(bb) - \varphi(a)\varphi(b)\varphi(a)\varphi(b)$. Analogously, all mixed moments can (at least in principle) be calculated by reducing them to alternating products of centered variables as in the definition of freeness. Thus the statements s_a, s_b are free, and each of them being semicircular determines all joint moments in s_a and s_b. This shows that s_a and s_b are not ordinary random variables but take values on some noncommutative algebra.

To apply the theory of free probability to RMT, we need to extend the definition of free to asymptotic freeness; that is, replacing the state functional φ by ϕ, where

$$\phi(\mathbf{A}) = \lim_{n\to\infty} \frac{1}{n} \mathrm{trE}(\mathbf{A}_n).$$

Since normalized traces of powers of a Hermitian matrix are the moments of the ESD of the matrix, free probability reveals important information on their LSD. It is shown that freeness of random matrices corresponds to independence and to distributions being invariant under orthogonal transformations. Formulas have been derived that express the LSD of sums and products of free random matrices in terms of their individual LSDs.

For an excellent introduction to free probability, see Biane [52] and Nica and Speicher [221].

Chapter 2
Wigner Matrices and Semicircular Law

A Wigner matrix is a symmetric (or Hermitian in the complex case) random matrix. Wigner matrices play an important role in nuclear physics and mathematical physics. The reader is referred to Mehta [212] for applications of Wigner matrices to these areas. Here we mention that they also have a strong statistical meaning. Consider the limit of a normalized Wishart matrix. Suppose that $\mathbf{x}_1, \cdots, \mathbf{x}_n$ are iid samples drawn from a p-dimensional multivariate normal population $N(\boldsymbol{\mu}, \mathbf{I}_p)$. Then, the sample covariance matrix is defined as

$$\mathbf{S}_n = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})',$$

where $\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$. When n tends to infinity, $\mathbf{S}_n \to \mathbf{I}_p$ and $\sqrt{n}\,(\mathbf{S}_n - \mathbf{I}_p) \to \sqrt{p}\mathbf{W}_p$. It can be seen that the entries above the main diagonal of $\sqrt{p}\mathbf{W}_p$ are iid $N(0,1)$ and the entries on the diagonal are iid $N(0,2)$. This matrix is called the (standard) Gaussian matrix or Wigner matrix.

A generalized definition of Wigner matrix only requires the matrix to be a Hermitian random matrix whose entries on or above the diagonal are independent. The study of spectral analysis of the large dimensional Wigner matrix dates back to Wigner's [295] famous **semicircular law**. He proved that the expected ESD of an $n \times n$ standard Gaussian matrix, normalized by $1/\sqrt{n}$, tends to the semicircular law F whose density is given by

$$F'(x) = \begin{cases} \frac{1}{2\pi}\sqrt{4 - x^2}, & \text{if } |x| \le 2, \\ 0, & \text{otherwise.} \end{cases} \tag{2.0.1}$$

This work has been extended in various aspects. Grenander [136] proved that $\|F^{\mathbf{W}_n} - F\| \to 0$ in probability. Further, this result was improved as in the sense of "almost sure" by Arnold [8, 7]. Later on, this result was further generalized, and it will be introduced in the following sections.

Z. Bai and J.W. Silverstein, *Spectral Analysis of Large Dimensional Random Matrices, Second Edition*, Springer Series in Statistics, DOI 10.1007/978-1-4419-0661-8_2,

2.1 Semicircular Law by the Moment Method

In order to apply the moment method (see Appendix B, Section B.1) to prove the convergence of the ESD of Wigner matrices to the semicircular distribution, we calculate the moments of the semicircular distribution and show that they satisfy the Carleman condition. In the remainder of this section, we will show the convergence of the ESD of the Wigner matrix by the moment method.

2.1.1 Moments of the Semicircular Law

Let β_k denote the k-th moment of the semicircular law. We have the following lemma.

Lemma 2.1. *For $k = 0, 1, 2, \cdots,$ we have*

$$\beta_{2k} = \frac{1}{k+1}\binom{2k}{k},$$
$$\beta_{2k+1} = 0.$$

Proof. Since the semicircular distribution is symmetric about 0, thus we have $\beta_{2k+1} = 0$. Also, we have

$$\begin{aligned}
\beta_{2k} &= \frac{1}{2\pi}\int_{-2}^{2} x^{2k}\sqrt{4-x^2}dx \\
&= \frac{1}{\pi}\int_{0}^{2} x^{2k}\sqrt{4-x^2}dx \\
&= \frac{2^{2k+1}}{\pi}\int_{0}^{1} y^{k-1/2}(1-y)^{1/2}dy \quad \text{(by setting } \ x = 2\sqrt{y}) \\
&= \frac{2^{2k+1}}{\pi}\frac{\Gamma(k+1/2)\Gamma(3/2)}{\Gamma(k+2)} = \frac{1}{k+1}\binom{2k}{k}.
\end{aligned}$$

2.1.2 Some Lemmas in Combinatorics

In order to calculate the limits of moments of the ESD of a Wigner matrix, we need some information from combinatorics. This is because the mean and variance of each empirical moment will be expressed as a sum of expectations of products of matrix entries, and we need to be able to systematically count the number of significant terms. To this end, we introduce some concepts from graph theory and establish some lemmas.

A graph is a triple (E, V, F), where E is the set of edges, V is the set of vertices, and F is a function, $F: E \mapsto V \times V$. If $F(e) = (v_1, v_2)$, the vertices v_1, v_2 are called the ends of the edge e, v_1 is the initial of e, and v_2 is the terminal of e. If $v_1 = v_2$, edge e is a loop. If two edges have the same set of ends, they are said to be coincident.

Let $\mathbf{i} = (i_1, \cdots, i_k)$ be a vector valued on $\{1, \cdots, n\}^k$. With the vector $\mathbf{i}$, we define a Γ-graph as follows. *Draw a horizontal line and plot the numbers* $i_1, \cdots, i_k$ *on it. Consider the distinct numbers as vertices, and draw* k *edges* e_j *from* i_j *to* i_{j+1}, $j = 1, \cdots, k$, *where* $i_{k+1} = i_1$ *by convention. Denote the number of distinct* i_j*'s by* t. Such a graph is called a $\Gamma(k,t)$-graph. An example of $\Gamma(6,4)$ is shown in Fig. 2.1.

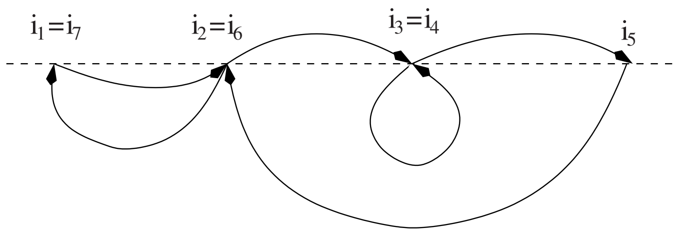

Fig. 2.1 A Γ-graph

By definition, a $\Gamma(k,t)$-graph starts from vertex i_1, and the k edges consecutively connect one after another and finally return to vertex i_1. That is, a $\Gamma(k,t)$-graph forms a cycle.

Two $\Gamma(k,t)$-graphs are said to be isomorphic if one can be converted to the other by a permutation of $(1, \cdots, n)$. By this definition, all Γ-graphs are classified into isomorphism classes.

We shall call the $\Gamma(k,t)$-graph canonical if it has the following properties:

1. Its vertex set is $V = \{1, \cdots, t\}$.
2. Its edge set is $E = \{e_1, \cdots, e_k\}$.
3. There is a function g from $\{1, 2, \cdots, k\}$ onto $\{1, 2, \cdots, t\}$ satisfying $g(1) = 1$ and $g(i) \le \max\{g(1), \cdots, g(i-1)\} + 1$ for $1 < i \le k$.
4. $F(e_i) = (g(i), g(i+1))$, for $i = 1, \cdots, k$, with convention $g(k+1) = g(1) = 1$.

It is easy to see that each isomorphism class contains one and only one canonical Γ-graph that is associated with a function g, and a general graph in this class can be defined by $F(e_j) = (i_{g(j)}, i_{g(j+1)})$. Therefore, we have the following lemma.

Lemma 2.2. *Each isomorphism class contains* $n(n-1)\cdots(n-t+1)$ $\Gamma(k,t)$ *graphs.*