Mathematics Monograph Series 15

Spectral Analysis of Large Dimensional Random Matrices

(Second Edition)

Zhidong Bai   Jack W. Silverstein

(大维随机矩阵的谱分析)

SCIENCE PRESS
Beijing

Responsible Editors: Zhao Yanchao, Lü Hong

Published by Science Press
16 Donghuangchenggen North Street
Beijing 100717, China

Printed in Beijing

ISBN 978-7-03-026777-1 (Peijing)

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
Cary, North Carolina, USA

Zhidong Bai
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
Cary, North Carolina, USA

Zhidong Bai
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 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

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

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 $H_0 : \operatorname{rank}(\boldsymbol{\Sigma}_q) = q$ against $H_1 : \operatorname{rank}(\boldsymbol{\Sigma}_q) > q$ is given by

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

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