Mathematics Monograph Series **2**

Spectral Analysis of Large Dimensional Random Matrices

Zhidong Bai   Jack W. Silverstein

（大维随机矩阵的谱分析）

SCIENCE PRESS
Beijing

Mathematics Monograph Series 2

Zhidong Bai Jack W. Silverstein

Spectral Analysis of Large Dimensional Random Matrices

(大维随机矩阵的谱分析)

SCIENCE PRESS
Beijing

Responsible Editor: Lü Hong

Published by Science Press
http://www.sciencep.com
16 Donghuangchenggen North Street
Beijing 100717, China

Printed in Beijing, 2006

ISBN 7-03-017766-5 (Beijing)

Preface

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 1940's and 50's. 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 semi-circular 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 Mathscinet database under keyword Random Matrix on 10 June 2005 (See Table 0.1.)

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

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

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 billions. In finance research, the number of different stocks can be as large as tens of thousands.

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

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.

With more than a half century's development of RMT, many different methodologies 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 either obtained 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 which we have not introduced.

The authors would like to express their appreciation to Professors Chen Mufa, Lin Qun, 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 2006

Contents

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 fixed dimension and those for 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 fold. This has enabled one to collect, store and analyze data sets of very high dimension. These computational developments have had strong impact on every branch of science. For example, R. A. 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 1970's; 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 1970's. It is difficult to imagine that these advanced areas in statistics would have gotten such deep stages of development if there were no such assistance from the present day computer.

Although modern computer technology helps us in so many aspects, it also brings a new and urgent task to the statisticians, that is, whether the classical limit theorems (i.e., those assuming 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 fixed dimension (classical limit theorems) and those for large dimension (large dimensional limit theorems). The problem turns out to be which kind of results is closer to reality? As argued in Huber (1973), some statisticians might say that five samples for each parameter in average are enough for using 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, or $p = 2\sqrt{n}$, or $p = 0.2n$. So, we have at least three different options to choose for an asymptotic setup. A natural question is then, which setup is the best choice among the three? Huber strongly suggested to study the situation of 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 4. Also, it is found that in linear regression analysis, if the covariates are random (or having 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 (MUltivariate 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 hundred-percent correct if the sample size is larger than 200. Why this paradox would 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 of 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 (1996) who theoretically proved that when testing the difference of means of two high dimensional populations, Dempster's (1959) non-exact 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 of dimension reduction were 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 3 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 3 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

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

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

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

where $\lambda_{n,j}$, $j = 1, \ldots, p$, are the eigenvalues of $\mathbf{S}_n$. 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

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

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 $\{\mathbf{S}_n\}$ [see Chapter 3], we will show that with probability one

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

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 which 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 in 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 been 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 1940's and early 1950's. In QM, the energy levels of a system are described by eigenvalues of an 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 attracts 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 (1990).

Since the late 1950's, research on the limiting spectral analysis of large dimensional random matrices has attracted considerable interest among mathematicians, probabilitists and statisticians. One pioneering work is the semicircular law for a Gaussian (or Wigner) matrix , due to Wigner, E., (Wigner (1955, 1958)). 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 (1967, 1971) and Grenander (1963) in various aspects. Bai and Yin (1988a) 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 (1967) and Pastur (1972, 1973), the asymptotic theory of spectral analysis of large dimensional sample covariance matrices was developed by many researchers including Bai, Yin, and Krishnaiah (1986), Grenander and Silverstein (1977), Jonsson (1982), Wachter (1978), Yin (1986), and Yin and Krishnaiah (1983). Also, Bai, Yin, and Krishnaiah (1986, 1987), Silverstein (1985a), Wachter (1980), Yin (1986), and Yin and Krishnaiah (1983) investigated the limiting spectral distribution of the multivariate F-matrix, or more generally, of products of random matrices. In the early 1980's, major contributions on the existence of LSD and their explicit forms for certain classes of random matrices were made. In recent years, research on RMT 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.

1.2.1 Spectral Analysis of Large Dimensional Random Matrices

Suppose **A** is an $m \times m$ matrix with eigenvalues λ_j, $j = 1, 2, \ldots, m$. 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 **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 **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), 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 give now a few examples.

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

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

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 $\sigma_1 \ge \sigma_2 \ge \dots \ge \sigma_p$. Then the test statistic for the hypothesis H_0 : rank$(\boldsymbol{\Sigma}_q) = q$ against H_1 : rank$(\boldsymbol{\Sigma}_q) > q$ is given by

$$\begin{aligned} T &= \frac{1}{p-q} \sum_{j=q+1}^{p} \sigma_j^2 - \left(\frac{1}{p-q} \sum_{j=q+1}^{p} \sigma_j \right)^2 \\ &= \frac{p}{p-q} \int_0^{\sigma_q} x^2 F^{\mathbf{S}}(dx) - \left(\frac{p}{p-q} \int_0^{\sigma_q} x F^{\mathbf{S}}(dx) \right)^2. \end{aligned}$$

1.2.2 Limits of Extreme Eigenvalues

In applications of the asymptotic theorems of spectral analysis of large dimensional random matrices, two important problems arose after the LSD was 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 (1980), 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 (1988) proved the same result under the existence of the 4th moment, and Bai, Silverstein, and Yin

(1988) proved that the existence of the 4th moment is also necessary for the existence of the limit. Bai and Yin (1988b) 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 was also found.

Comparing 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 (1983), 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 dimension to the degrees of freedom is less than 1/2. Silverstein (1984) modified the work to allowing the ratio less than 1. Silverstein (1985b) further proved that with probability one, the smallest eigenvalue of a Wishart matrix tends to the lower bound of the LSD when the ratio of 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, it cannot be extended to the general case. The most current contribution was made in Bai and Yin (1993) in which it is proved that under the existence of the fourth moment of the underlying distribution, the smallest eigenvalue (when $p \leq n$) or the $p - n + 1$st smallest eigenvalue (when $p > n$) tends to $a(y) = \sigma^2(1 - \sqrt{y})^2$, where $y = \lim(p/n) \in (0, \infty)$. Comparing 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 ESD

The second problem, the convergence rate of the spectral distributions of large dimensional random matrices, is of practical interest, but has 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 (1993a, b), 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 $O(n^{-1/4})$, 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 apart away from one, and to be $O(n^{-5/48})$, if the ratio is close to 1.

1.2.4 Circular Law

The most perplexing problem is the so-called circular law which conjectures that the spectral distribution of a non-symmetric random matrix, after suitable normalization, tends to the uniform distribution over the unit disc in the complex plane. The difficulty exists in that two most important tools used for symmetric matrices do not apply for non-symmetric matrices. Furthermore, certain truncation and centralization techniques cannot be used. The first known result was given in Mehta (1967) and in an unpublished paper of Silverstein (1984) which was reported in Hwang (1986). 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 which was found by Ginibre (1965). The first attempt to prove this conjecture under some general conditions was made in Girko (1984a, b). However, his proofs have puzzled many who attempt to understand, without success, Girko's arguments. Recently, Edelman (1995) 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 $4+\varepsilon$ moment and some smooth conditions, Bai (1997) proved the strong version of the circular law.

1.2.5 CLT of Linear Spectral Statisticslinear spectral statistics

As mentioned above, functionals of the ESD of RM's 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 (2004) the normalization has been 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 D. Jonsson (1982) in which $f(t) = t^r$ and F_n is the ESD of normalized standard Wishart matrix. Further work was done by Johansson, K. (1998), Bai and Silverstein (2004), Bai and Yao (2004), Sinai and Soshnikov (1998), among others.