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

(O-2562.0101)

Spectral Analysis of Large Dimensional Random Matrices

The aim of the book is to introduce basic concepts, main results, and widely applied mathematical tools in the spectral analysis of large dimensional random matrices. In it we will introduce many of the fundamental results, such as the semicircular law of Wigner matrices, the Marchenko-Pastur law, the limiting spectral distribution of the multivariate F matrix, limits of extremal eigenvalues, spectrum separation theorems, convergence rates of empirical spectral distributions, central limit theorems of linear spectral statistics and the partial solution of the famous circular law. While deriving the main results, the book will simultaneously emphasize the ideas and methodologies of the fundamental mathematical tools, among them being: truncation techniques, matrix transformations, moment convergence theorems, and the Stieltjes transform. Thus, its treatment is especially fitting to the needs of mathematics and statistic graduate students, and beginning researchers, who can learn the basic methodologies and ideas to solve problems in this area. It may also serve as a detailed handbook on results of large dimensional random matrices for practical users.

ISBN 7-03-017766-5
9 787030 177667 >

ISBN 7-03-017766-5
定 价：80.00 元

Mathematics Monograph Series

2

Spectral Analysis of Large Dimensional Random Matrices

in

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)

To my great teacher Professor Yongquan Yin
my wife Xicun Dan
and my sons Li and Steve Gang

— Zhidong Bai

To my children Hila and Idan

— Jack W. Silverstein

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