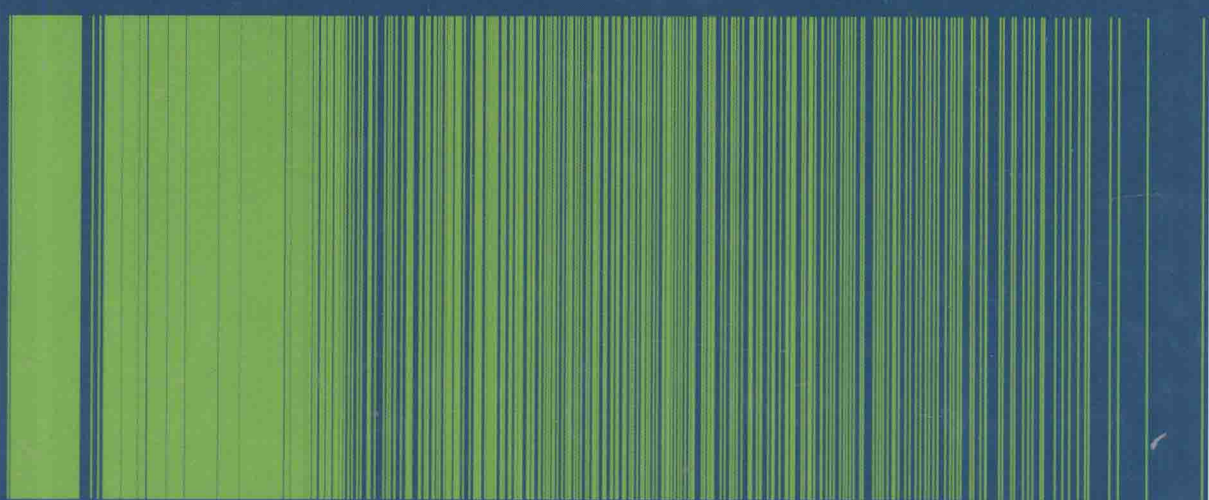
Cambridge Series in Statistical and Probabilistic Mathematics

Large Sample Covariance Matrices and High-Dimensional Data Analysis

Jianfeng Yao
Shurong Zheng
Zhidong Bai



Large Sample Covariance Matrices and High-Dimensional Data Analysis

Jianfeng Yao
The University of Hong Kong

Shurong Zheng
Northeast Normal University, China

Zhidong Bai
Northeast Normal University, China

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

32 Avenue of the Americas, New York, NY 10013-2473, USA

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107065178

First published 2015

Printed in Great Britain by Clays Ltd, St Ives plc

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging in Publication Data
Yao, Jianfeng.
Large sample covariance matrices and high-dimensional data analysis / Jianfeng Yao, University of Hong Kong, Shurong Zheng; Northeast Normal University, China, Zhidong Bai, Northeast Normal University, China
pages cm. – (Cambridge series in statistical and probabilistic mathematics)
Includes bibliographical references and index.
ISBN 978-1-107-06517-8 (hardback)
1. Analysis of covariance. 2. Multivariate analysis. 3. Statistics. I. Bai, Zhidong.
II. Zheng, Shurong. III. Title.
QA279.Y366 2015
519.5′38–dc23 2014044911

ISBN 978-1-107-06517-8 Hardback

Large Sample Covariance Matrices and High-Dimensional Data Analysis

High-dimensional data appear in many fields, and their analysis has become increasingly important in modern statistics. However, it has long been observed that several well-known methods in multivariate analysis become inefficient, or even misleading, when the data dimension p is larger than, say, several tens. A seminal example is the well-known inefficiency of Hotelling's T^2-test in such cases. This example shows that classical large sample limits yield poor approximations for high-dimensional data; statisticians must seek new limiting theorems in these instances. Thus, the theory of random matrices (RMT) serves as a much-needed and welcome alternative framework. Based on the authors' own research, this book provides a firsthand introduction to new high-dimensional statistical methods derived from RMT. The book begins with a detailed introduction to useful tools from RMT and then presents a series of high-dimensional problems with solutions provided by RMT methods.

JIANFENG YAO has rich research experience on random matrix theory and its applications to high-dimensional statistics. In recent years, he has published many authoritative papers in these areas and organised several international workshops on related topics.

SHURONG ZHENG is author of several influential results in random matrix theory including a widely used central limit theorem for eigenvalue statistics of a random Fisher matrix. She has also developed important applications of the inference theory presented in the book to real-life high-dimensional statistics.

ZHIDONG BAI is a world-leading expert in random matrix theory and high-dimensional statistics. He has published more than 200 research papers and several specialized monographs, including *Spectral Analysis of Large Dimensional Random Matrices* (with J. W. Silverstein), for which he won the Natural Science Award of China (Second Class in 2012).

CAMBRIDGE SERIES IN STATISTICAL AND PROBABILISTIC MATHEMATICS

This series of high-quality upper-division textbooks and expository monographs covers all aspects of stochastic applicable mathematics. The topics range from pure and applied statistics to probability theory, operations research, optimization, and mathematical programming. The books contain clear presentations of new developments in the field and also of the state of the art in classical methods. While emphasizing rigorous treatment of theoretical methods, the books also contain applications and discussions of new techniques made possible by advances in computational practice.

A complete list of books in the series can be found at www.cambridge.org/statistics. Recent titles include the following:

14. *Statistical Analysis of Stochastic Processes in Time*, by J. K. Lindsey
15. *Measure Theory and Filtering*, by Lakhdar Aggoun and Robert Elliott
16. *Essentials of Statistical Inference*, by G. A. Young and R. L. Smith
17. *Elements of Distribution Theory*, by Thomas A. Severini
18. *Statistical Mechanics of Disordered Systems,* by Anton Bovier
19. *The Coordinate-Free Approach to Linear Models*, by Michael J. Wichura
20. *Random Graph Dynamics*, by Rick Durrett
21. *Networks*, by Peter Whittle
22. *Saddlepoint Approximations with Applications*, by Ronald W. Butler
23. *Applied Asymptotics*, by A. R. Brazzale, A. C. Davison and N. Reid
24. *Random Networks for Communication*, by Massimo Franceschetti and Ronald Meester
25. *Design of Comparative Experiments*, by R. A. Bailey
26. *Symmetry Studies*, by Marlos A. G. Viana
27. *Model Selection and Model Averaging*, by Gerda Claeskens and Nils Lid Hjort
28. *Bayesian Nonparametrics*, edited by Nils Lid Hjort et al.
29. *From Finite Sample to Asymptotic Methods in Statistics*, by Pranab K. Sen, Julio M. Singer and Antonio C. Pedrosa de Lima
30. *Brownian Motion*, by Peter Mörters and Yuval Peres
31. *Probability (Fourth Edition)*, by Rick Durrett
32. *Analysis of Multivariate and High-Dimensional Data*, by Inge Koch
33. *Stochastic Processes*, by Richard F. Bass
34. *Regression for Categorical Data*, by Gerhard Tutz
35. *Exercises in Probability (Second Edition)*, by Loïc Chaumont and Marc Yor
36. *Statistical Principles for the Design of Experiments*, by R. Mead, S. G. Gilmour and A. Mead
37. *Quantum Stochastics*, by Mou-Hsiung Chang
38. *Nonparametric Estimation under Shape Constraints*, by Piet Groeneboom and Geurt Jungbloed

This book is dedicated to Xavier Guyon and Yongquan Yin.
We also dedicate this book to our families:

Alice, Céline, Jérémy, Thaïs and Yan,
Yuanning and Guanghou,
Xicun, Li, Gang, Yongji, Yonglin and Yongbin.

Notation

$\overset{\mathscr{D}}{=}$	equality in distribution
$\xrightarrow{\mathscr{D}}$	convergence in distribution
$\xrightarrow{\text{a.s.}}$	almost sure convergence
$\xrightarrow{\mathscr{P}}$	convergence in probability
CLT	central limit theorem
δ_{jk}	Kronecker symbol: 1/0 for $j = k / j \neq k$
δ_a	Dirac mass at a
$\mathbf{e}_j$	jth vector of a canonical basis
ESD	empirical spectral distribution
Γ_μ	support set of a finite measure μ
$I_{(\cdot)}$	indicator function
$\mathbf{I}_p$	p-dimensional identity matrix
LSD	limiting spectral distribution
MP	Marčenko-Pastur
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$o_P(1)$, $O_P(1)$, $o_{\text{a.s}}(1)$, $O_{\text{a.s.}}(1)$	stochastic order symbols
PSD	population spectral distribution
$\mathbf{u}$, $\mathbf{X}$, $\boldsymbol{\Sigma}$, etc.	vectors and matrices are boldfaced

Preface

Dempster (1958, 1960) proposed a non-exact test for the two-sample significance test when the dimension of data is larger than the degrees of freedom. He raised the question of what statisticians should do if traditional multivariate statistical theory does not apply when the dimension of data is too large. Later, Bai and Saranadasa (1996) found that even when traditional approaches can be applied, they are much less powerful than the non-exact test when the dimension of data is large. This raised another question of how classical multivariate statistical procedures could be adapted and improved when the data dimension is large. These problems have attracted considerable attention since the middle of the first decade of this century. Efforts towards solving these problems have been made along two directions: the first is to propose special statistical procedures to solve ad hoc large-dimensional statistical problems where traditional multivariate statistical procedures are inapplicable or perform poorly, for some specific large-dimensional hypotheses. The family of various non-exact tests follows this approach. The second direction, following the work of Bai et al. (2009a), is to make systematic corrections to the classical multivariate statistical procedures so that the effect of large dimension is overcome. This goal is achieved by employing new and powerful asymptotic tools borrowed from the theory of random matrices, such as the central limit theorems in Bai and Silverstein (2004) and Zheng (2012).

Recently, research along these two directions has become very active in response to an increasingly important need for analysis of massive and large-dimensional data. Indeed, such "big data" are nowadays routinely collected owing to rapid advances in computer-based or web-based commerce and data-collection technology.

To accommodate such need, this monograph collects existing results along the aforementioned second direction of large-dimensional data analysis. In Chapters 2 and 3, the core of fundamental results from random matrix theory about sample covariance matrices and random Fisher matrices is presented in detail. Chapters 4–12 collect large-dimensional statistical problems in which the classical large sample methods fail and the new asymptotic methods, based on the fundamental results of the preceding chapters, provide a valuable remedy. As the employed statistical and mathematical tools are quite new and technically demanding, our objective is to describe the state of the art through an accessible introduction to these new statistical tools. It is assumed that the reader is familiar with the usual theory of mathematical statistics, especially methods dealing with multivariate normal samples. Other prerequisites include knowledge of elementary matrix algebra and limit theory (the law of large numbers and the central limit theorem) for independent and identically distributed samples. A special prerequisite is some familiarity with contour integration; however, a detailed appendix on this topic has been included.

Readers familiar with Anderson's (2003) textbook *An Introduction to Multivariate Statistical Analysis* will easily recognise that our introduction to classical multivariate statistical methods, such as in Chapters 4, 7, 8 and 9, follows that textbook closely. We are deeply grateful to Anderson's phenomenal text, which has been a constant help during the preparation of this book.

This text has also benefited over the years from numerous collaborations with our colleagues and research students. We particularly thank the following individuals, whose joint research work with us has greatly contributed to the material presented in the book: Jiaqi Chen, Bernard Delyon, Xue Ding, Dandan Jiang, Hua Li, Weiming Li, Zhaoyuan Li, Huixia Liu, Guangming Pan, Damien Passemier, Yingli Qin, Hewa Saranadasa, Jack Silverstein, Qinwen Wang and Wing-Keung Wong.

Finally, two of us owe a debt of gratitude to Zhidong Bai: he has been for years a constant inspiration to us. This text would never have been possible without his outstanding leadership. We are particularly proud of the completion of the text in the year of his 70th birthday.

Contents

1

Introduction

1.1 Large-Dimensional Data and New Asymptotic Statistics

In a multivariate analysis problem, we are given a sample $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ of random observations of dimension p. Statistical methods, such as principal component analysis, have been developed since the beginning of the 20th century. When the observations are Gaussian, some nonasymptotic methods exist, such as Student's test, Fisher's test, or the analysis of variance. However, in most applications, observations are non-Gaussian, at least in part, so that nonasymptotic results become hard to obtain and statistical methods are built using limiting theorems on model statistics.

Most of these asymptotic results are derived under the assumption that the data dimension p is fixed while the sample size n tends to infinity (large sample theory). This theory had been adopted by most practitioners until very recently, when they were faced with a new challenge: the analysis of large dimensional data.

Large-dimensional data appear in various fields for different reasons. In finance, as a consequence of the generalisation of Internet and electronic commerce supported by the exponentially increasing power of computing, online data from markets around the world are accumulated on a giga-octet basis every day. In genetic experiments, such as micro-arrays, it becomes possible to record the expression of several thousand of genes from a single tissue. Table 1.1 displays some typical data dimensions and sample sizes. We can see from this table that the data dimension p is far from the "usual" situations where p is commonly less than 10. We refer to this new type of data as *large-dimensional data*.

It has been observed for a long time that several well-known methods in multivariate analysis become inefficient or even misleading when the data dimension p is not as small as, say, several tens. A seminal example was provided by Dempster in 1958, when he established the inefficiency of Hotelling's T^2 in such cases and provided a remedy (named a non-exact test). However, by that time, no statistician was able to discover the fundamental reasons for such a breakdown in the well-established methods.

To deal with such large-dimensional data, a new area in asymptotic statistics has been developed where the data dimension p is no longer fixed but tends to infinity *together* with the sample size n. We call this scheme *large-dimensional asymptotics*. For multivariate analysis, the problem thus turns out to be which one of the large sample scheme and the large-dimensional scheme is closer to reality. As Huber (1973) argued, some statisticians might say that five samples for each parameter on average is enough to use large sample 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

Table 1.1. *Examples of large-dimensional data*

	Data dimension p	Sample size n	$y = p/n$
Portfolio	~ 50	500	0.1
Climate survey	320	600	0.21
Speech analysis	$a \times 10^2$	$b \times 10^2$	~ 1
ORL face database	1440	320	4.5
Micro-arrays	1000	100	10

(large sample asymptotics), $p = 2\sqrt{n}$, or $p = 0.2n$ (large-dimensional asymptotics). So, we have at least three different options among which to choose for an asymptotic setup. A natural question then, is, which setup is the best choice among the three? Huber strongly suggested studying 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 four. Also, it is found that in linear regression analysis, if the covariates are random (or have measurement errors) and the number of covariates is larger than six, the behaviour of the estimates departs far from the theoretical values, unless the sample size is very large. In signal processing, when the number of signals is 2 or 3 and the number of sensors is more than 10, the traditional multivariate signal classification (music) 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 percent correct if the sample size is larger than 200. Why would this paradox occur? 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 proportionally to p^2, while the number $(2np)$ of observations increases proportionally 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 (1996), who theoretically proved that when testing the difference of means of two high-dimensional populations, the Dempster (1958) 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 for 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 three principal components so that more than 80 percent 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 large dimensional problem. If, again, three principal components only are selected, 90 percent or even more of the information carried in the original data set could be lost. Now, let us consider another example.