An Introduction to Large Sample covariance matrices and Their Applications

(June 2018)

Jianfeng Yao
Department of Statistics and Actuarial Science
The University of Hong Kong.

These notes are designed for a short course at the Insititute of Statistics and Big Data, Renmin University, in July 2018.

The main material in these notes is extracted from the book *Large Sample Covariance Matrices and High-dimensional Data Analysis* published by Cambridge University Press in 2015 (ISBN: 9781107065178). The author is grateful to his co-authors Prof. Zhidong Bai and Prof. Shurong Zheng for their significant contribution to the material presented here.

Contents

Notation *page* v

1 **Introduction** 1

1.1 Large dimensional data and new asymptotic statistics 1

1.2 Random matrix theory 3

1.3 Eigenvalue statistics of large sample covariance matrices 4

1.4 Organisation of these notes 5

2 **Limiting spectral distributions** 7

2.1 Introduction 7

2.2 Fundamental tools 8

2.3 Marčenko-Pastur distributions 10

2.4 Generalised Marčenko-Pastur distributions 16

2.5 LSD for random Fisher matrices 22

3 **CLT for linear spectral statistics** 30

3.1 Introduction 30

3.2 CLT for linear spectral statistics of a sample covariance matrix 31

3.3 Bai and Silverstein's CLT 39

3.4 CLT for linear spectral statistics of random Fisher matrices 40

3.5 The substitution principle 44

4 **The generalised variance and multiple correlation coefficient** 47

4.1 Introduction 47

4.2 The generalised variance 47

4.3 The multiple correlation coefficient 52

5 **Testing hypotheses of equality of covariance matrices** 58

5.1 Introduction 58

5.2 Testing equality between several covariance matrices 58

5.3 Testing equality of several multivariate normal distributions 60

5.4 The sphericity test 62

5.5 Testing equality about one covariance matrix 64

5.6 Testing hypotheses of equality of large covariance matrices 65

5.7 Large-dimensional sphericity test 74

6 **Large-dimensional spiked population models** 86

6.1 Introduction 86

6.2 Limits of spiked sample eigenvalues 88
6.3 Limits of spiked sample eigenvectors 94
6.4 Central limit theorem for spiked sample eigenvalues 96
6.5 Estimation of the values of spike eigenvalues 109
6.6 Estimation of the number of spike eigenvalues 111
6.7 Estimation of the noise variance 122
Appendix A **Curvilinear integrals** 129
Appendix B **Eigenvalue inequalities** 136

Bibliography 139
Index 143

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

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 Components 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 of 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 has been adopted by most of practitioner until very recently when they are faced with a new challenge, the analysis of large dimensional data.

Large dimensional data appear in various fields due to different reasons. In finance, as a consequence of the generalisation of Internet and electronic commerce supported by an exponentially increasing power of computing, on-line data from markets around the world are accumulated in a giga-octets basis every day. In genetic experiments such as micro-arrays, it becomes possible to record the expression of several thousands 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 "usual" situations where p is commonly seen smaller then 10. We refer this new type of data as *large dimensional data.*

It has been observed since 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

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	$a \cdot 10^2$	$b \cdot 10^2$	~ 1
ORL face data base	1440	320	4.5
micro-arrays	1000	100	10

as say several tens. A seminar example is provided by Dempster in 1958 where he established the inefficiency of Hotellings' T^2 in such cases and provided a remedy (named as a non-exact test). However, by that time no statistician was able to discover the fundamental reasons for such break-down of 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 more 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 argued in Huber (1973), some statisticians might say that five samples for each parameter in average are enough for using 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 (large sample asymptotics), $p = 2\sqrt{n}$ or $p = 0.2n$ (large dimensional asymptotics). 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 four. 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 behaviour of the estimates departs far away from the theoretical 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 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 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 (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 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 large dimensional problem. If again 3 principal components only are selected, 90% or even more of the information carried in the original data set could be lost. Now, let us consider another example.

Example 1.1 Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be a sample from p-dimensional Gaussian distribution $\mathcal{N}_p(0, \mathbf{I}_p)$ with mean zero and unit covariance matrix. The corresponding sample covariance matrix is

$$\mathbf{S}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^* .$$

An important statistic in multivariate analysis is

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

where $\{\lambda_{n,j}\}_{1\le j\le p}$ are the eigenvalues of $\mathbf{S}_n$. When p is fixed, $\lambda_{n,j}{\to}1$ almost surely as $n{\to}\infty$ and thus $T_n{\to}0$. Further, by taking a Taylor expansion of $\log(1 + x)$, one can show that

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

for any fixed p. This suggests the possibility that T_n remains asymptotically Gaussian for large p 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 2], it is readily seen that almost surely,

$$\frac{1}{p}T_n{\to}\int_{a(y)}^{b(y)} \frac{\log x}{2\pi yx}\left[\{b(y) - x\}\{x - a(y)\}\right]^{1/2} dx = \frac{y-1}{y}\log(1 - y) - 1 \equiv d(y) < 0 ,$$

where $a(y) = (1 - \sqrt{y})^2$ and $b(y) = (1 + \sqrt{y})^2$ (details of this calculation of integral are given in Example 2.11). This shows that almost surely

$$\sqrt{\frac{n}{p}}T_n \simeq d(y)\sqrt{np}{\to} - \infty.$$

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

These examples show that the classical large sample limits are no longer suitable for dealing with large dimensional data analysis. Statisticians must seek out new limiting theorems to deal with large dimensional statistical problems. In this context, the theory of random matrices (RMT) proves to be a powerful tool for achieving this goal.

1.2 Random matrix theory

RMT traces back to the development of quantum mechanics in the 1940s and the early 1950s. In this field, 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 discretisation, amounting to a truncation, keeping only the part of the Hilbert space that is important to the problem under consideration. Thus **A** becomes a finite but large dimensional random linear operator, i.e. a large dimensional random matrix. Hence, the limiting behaviour of large dimensional random matrices attracts special interest among experts in quantum mechanics and many limiting laws were discovered during that time. For a more detailed review on applications of RMT in quantum mechanics and other related areas in physics, the reader is referred to the Book *Random Matrices* by Mehta (2004).

Since the late 1950s, research on the limiting spectral properties 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 , due to E. Wigner (1955; 1958). He proved that the expected spectral distribution of a large dimensional Wigner matrix tends to the semicircular law. This work was later generalised by Arnold (1967, 1971) and Grenander (1963) in various aspects. On the another direction related to the class of Gaussian Wishart matrices, or more generally, the class of sample covariance matrices, the breakthrough work was done in Marčenko and Pastur (1967) and Pastur (1972, 1973) where the authors discovered the Marčenko-Pastur law under fairly general conditions. The asymptotic theory of spectral analysis of large dimensional sample covariance matrices was later developed by many researchers including Bai et al. (1986), Grenander and Silverstein (1977), Jonsson (1982), Wachter (1978), Yin (1986), and Yin and Krishnaiah (1983). Also, Bai et al. (1986, 1987), Silverstein (1985), Wachter (1980), Yin (1986), and Yin and Krishnaiah (1983) investigated the limiting spectral distribution of the multivariate Fisher matrix, or more generally, of products of random matrices (a random Fisher matrix is the product of a sample covariance matrix by the inverse of another independent sample covariance matrix). In the early 1980s, major contributions on the existence of limiting spectral distributions and their explicit forms for certain classes of random matrices were made. In particular, Bai and Yin (1988) proved that the spectral distribution of a sample covariance matrix (suitably normalised) tends to the semicircular law when the dimension is relatively smaller than the sample size. In recent years, research on RMT is turning toward the second order limiting theorems, such as the central limit theorem for linear spectral statistics, the limiting distributions of spectral spacings and extreme eigenvalues.

1.3 Eigenvalue statistics of large sample covariance matrices

This book is about the theory of large sample covariance matrices and their applications to high-dimensional statistics. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be a sample of random observations of dimension p. The population covariance matrix is denoted by $\mathbf{\Sigma} = \text{cov}(\mathbf{x}_i)$. The corresponding *sample covariance matrix* is defined as

$$\mathbf{S}_n = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^*, \tag{1.1}$$

where $\overline{\mathbf{x}} = n^{-1} \sum_i \mathbf{x}_i$ denotes the *sample mean*. Almost all statistical methods in multivariate analysis rely on this sample covariance matrix: principle component analysis, canonical correlation analysis, multivariate regressions, one-sample or two-sample hypothesis testing, factor analysis etc.

A striking fact in multivariate analysis or large dimensional statistics is that many important statistics are function of the eigenvalues of sample covariance matrices. The statistic T_n in Example 1.1 is of this type and here is yet another example.

Example 1.2 Let the covariance matrix of a population have the form $\mathbf{\Sigma} = \mathbf{\Sigma}_q + \sigma^2\mathbf{I}$, where $\mathbf{\Sigma}$ is $p \times p$ and $\mathbf{\Sigma}_q$ has rank q $(q < p)$. Suppose $\mathbf{S}_n$ is the sample covariance matrix based on a sample of size n drawn from the population. Denote the eigenvalues of $\mathbf{S}_n$ by $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Then the test statistic for the hypothesis H_0: rank$(\mathbf{\Sigma}_q) = q$ against H_1: rank$(\mathbf{\Sigma}_q) > q$ is given by

$$Q_n = \frac{1}{p-q} \sum_{j=q+1}^{p} \lambda_j^2 - \left(\frac{1}{p-q} \sum_{j=q+1}^{p} \lambda_j \right)^2 .$$

In other words, the test statistic Q_n is the variance of the $p-q$ smallest eigenvalues of $\mathbf{S}_n$.

Therefore, understanding the asymptotic properties of eigenvalue statistics such as T_n and Q_n above has a paramount importance in data analysis when the dimension p is getting large with respect to the sample size. The spectral analysis of large dimensional sample covariance matrices from RMT provides powerful tools for the study of such eigenvalue statistics. For instance, the Marčenko-Pastur law describe the global behaviour of the p eigenvalues of a sample covariance matrix so that point-wise limits of eigenvalue statistics are determined by integrals of appropriate functions with respect to the Marčenko-Pastur law, see Example 1.1 for the case of T_n. Moreover, fluctuations of these eigenvalue statistics are described by central limit theorems which are found in Bai and Silverstein (2004) and in Zheng (2012). Similarly to the case of classical large sample theory, such CLTs constitute the corner-stones of statistical inference with large dimensional data.

1.4 Organisation of these notes

In Chapters 2 and 3, the core of fundamental results from RMT regarding sample covariance matrices and random Fisher matrices is presented in details. These results are selected in such a way that they are applied and used in the subsequent chapters of the book. More specifically, Chapter 2 introduces the limiting spectral distributions of general sample covariance matrices, namely the Marčenko-Pastur distributions, and the limiting spectral distributions of random Fisher matrices. Detailed examples of both limits are also provided. In Chapter 3, the two fundamental CLTs from Bai and Silverstein (2004) and Zheng (2012) are presented in details. Simple application examples of these CLTs are given. We also introduce a *substitution principle* that deals with the effect in the CLTs induced by the use of adjusted sample sizes $n_i - 1$ in place of the (raw) sample sizes n_i in the definition of sample covariance matrices and Fisher matrices.

The Chapters 4, 5 and 6 develop several large dimensional statistical problems where the classical large sample methods fail and the new asymptotic methods from the above RMT provide a valuable remedy. Topics in Chapter 4 and Chapter 5 are classical topics in multivariate analysis; they are here re-analysed under the large-dimensional scheme. The last chapter treats a modern topic in large-dimensional statistics.

An appendix is included to introduce the basics on contour integration. The reason is that in the CLT's developed in Chapter 3 for linear spectral statistics of sample covariance matrices and of random Fisher matrices, the mean and covariance functions of the limiting Gaussian distributions are expressed in terms of contour integrals, and explicit calculations of such contour integrals frequently appear at various places of this book.

Notes

On the interplay between the random matrix theory and large-dimensional statistics, supplementary information can be found in the excellent introductory papers Bai (2005) Johnstone (2007) and Johnstone and Titterington (2009).

2

Limiting spectral distributions

2.1 Introduction

Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be a sample of random observations of dimension p. The *sample covariance matrix* is defined as

$$\mathbf{S}_n = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^* = \frac{1}{n-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^* - \frac{n}{n-1}\overline{\mathbf{x}\mathbf{x}}^*, \tag{2.1}$$

where $\overline{\mathbf{x}} = n^{-1}\sum_i \mathbf{x}_i$ denotes the *sample mean*. Many of traditional multivariate statistics are functions of the eigenvalues $\{\lambda_k\}$ of the sample covariance matrix $\mathbf{S}_n$. In the most basic form, such statistics can be written as

$$T_n = \frac{1}{p}\sum_{k=1}^{p}\varphi(\lambda_k)\,, \tag{2.2}$$

for some specific function φ. Such statistic is called a *linear spectral statistic* of the sample covariance matrix $\mathbf{S}_n$. For example, the so-called *generalised variance* discussed later in Chapter 4, see Eq.(4.1) is

$$T_n = \frac{1}{p}\log|\mathbf{S}_n| = \frac{1}{p}\sum_{k=1}^{p}\log(\lambda_k).$$

So this particular T_n is a linear spectral statistic of the sample covariance matrix $\mathbf{S}_n$ with "test function" $\varphi(x) = \log(x)$.

In two-sample multivariate analysis with say an $\mathbf{x}$-sample and an $\mathbf{y}$-sample, interesting statistics will still be of the previous form in (2.2), where however the eigenvalues $\{\lambda_k\}$ will be those of the so-called *Fisher matrix* $\mathbf{F}_n$. Notice that each of the two examples has a corresponding sample covariance matrix, say $\mathbf{S}_\mathbf{x}$ and $\mathbf{S}_\mathbf{y}$. The Fisher matrix associated to these samples is the quotient of the two sample matrices, namely $\mathbf{F}_n = \mathbf{S}_\mathbf{x}\mathbf{S}_\mathbf{y}^{-1}$ (assuming the later is invertible).

Linear spectral statistics of sample covariance matrices or Fisher matrices are at the heart of the new statistical tools developed in this book. In this chapter and the next Chapter 3, we introduce the theoretical backgrounds on these statistics. More specifically, this chapter deals with the first order limits of such statistics, namely to answer the question:

When and how T_n should converge to some limiting value ℓ as both the dimension p and the sample size grow to infinity?

Clearly, the question should relate to the "joint limit" of the p eigenvalues $\{\lambda_k\}$. The formal concepts to deal with the question are called the *empirical spectral distributions* and *limiting spectral distributions*. In this chapter, these distributions for the sample covariance matrix $\mathbf{S}_n$ and the two-sample Fisher matrix $\mathbf{F}_n$ are introduced.

2.2 Fundamental tools

This section introduces some fundamental concepts and tools used throughout the book.

2.2.1 Empirical and limiting spectral distributions

Let $\mathcal{M}_p(\mathbb{C})$ be the set of $p \times p$ matrices with complex-valued elements.

Definition 2.1 Let $\mathbf{A} \in \mathcal{M}_p(\mathbb{C})$ and $\{\lambda_j\}_{1\le j\le p}$, its *empirical spectral distribution* (ESD) is

$$F^{\mathbf{A}} = \frac{1}{p}\sum_{j=1}^{p} \delta_{\lambda_j} \ ,$$

where δ_a denotes the Dirac mass at a point a.

In general, the ESD $F^{\mathbf{A}}$ is a probability measure on $\mathbb{C}$; it has support in $\mathbb{R}$ (resp. on $\mathbb{R}_+$) if $\mathbf{A}$ is Hermitian (resp. nonnegative definite Hermitian). For example, the two-dimensional rotation

$$\mathbf{A} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

has eigenvalues $\pm i$ so that $F^{\mathbf{A}} = \frac{1}{2}(\delta_{\{i\}} + \delta_{\{-i\}})$ is a measure on $\mathbb{C}$, while the symmetry

$$\mathbf{B} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

has eigenvalues ± 1 so that $F^{\mathbf{B}} = \frac{1}{2}(\delta_{\{1\}} + \delta_{\{-1\}})$ has support on $\mathbb{R}$. In this book, we are mainly concerned by covariance matrices. Since there are Hermitian and nonnegative definite, the corresponding ESD's will have support on $\mathbb{R}_+$.

Definition 2.2 Let $\{\mathbf{A}_n\}_{n\ge 1}$ be a sequence from $\mathcal{M}_p(\mathbb{C})$. If the sequence of corresponding ESD's $\{F^{\mathbf{A}_n}\}_{n\ge 1}$ vaguely converges to a (possibly defective) measure F, we call F the *limiting spectral distribution* (LSD) of the sequence of matrices $\{\mathbf{A}_n\}$.

The above vague convergence means that for any continuous and compactly supported function φ, $F^{\mathbf{A}_n}(\varphi) \to F(\varphi)$ as $n \to \infty$. It is well-known that if the LSD F is indeed non defective, i.e. $\int F(dx) = 1$, the above vague convergence turns into the stronger (usual) weak convergence, i.e. $F^{\mathbf{A}_n}(\varphi) \to F(\varphi)$ for any continuous and bounded function φ.

When dealing with a sequence of sample covariance matrices $\{\mathbf{S}_n\}$, their eigenvalues are random variables and the corresponding ESD's $\{F^{\mathbf{S}_n}\}$ are random probability measures on $\mathbb{R}_+$. A fundamental question in random matrix theory is about whether the sequence $\{F^{\mathbf{S}_n}\}$ has a limit (in probability or almost surely).

2.2.2 Stieltjes transform

The eigenvalues of a matrix are continuous functions of entries of the matrix. But these functions have no closed forms when the dimension of the matrix is larger than four. So special methods are needed for their study. There are three important methods employed in this area, moment method, Stieltjes transform and orthogonal polynomial decomposition of the exact density of eigenvalues. For the sake of our exposition, we concentrate on the Stieltjes transform method which is indeed widely used in the literature of large dimensional statistics.

We denote by Γ_μ the support of a finite measure μ on $\mathbb{R}$. Let

$$\mathbb{C}^+ := \{z \in \mathbb{C} \ : \ \Im(z) > 0\}$$

be the (open) upper half complex plan with positive imaginary part.

Definition 2.3 Let μ be a finite measure on the real line. Its Stieltjes transform (also called Cauchy transform in the literature) is defined as

$$s_\mu(z) = \int \frac{1}{x-z}\mu(dx) , \quad z \in \mathbb{C} \setminus \Gamma_\mu .$$

The results of this section are given without proofs; they can be found in textbooks such as Kreĭn and Nudel′man (1977).

Proposition 2.4 *The Stieltjes transform has the following properties:*

(i) *s_μ is holomorphic on $\mathbb{C} \setminus \Gamma_\mu$;*
(ii) *$z \in \mathbb{C}^+$ if and only if $s_\mu(z) \in \mathbb{C}^+$;*
(iii) *If $\Gamma_\mu \subset \mathbb{R}_+$ and $z \in \mathbb{C}^+$, then $zs_\mu(z) \in \mathbb{C}^+$;*
(iv) $|s_\mu(z)| \le \dfrac{\mu(1)}{dist(z, \Gamma_\mu) \vee |\Im(z)|}$.

The next result is an inversion result.

Proposition 2.5 *The mass $\mu(1)$ can be recovered through the formula*

$$\mu(1) = \lim_{v\to\infty} -ivs_\mu(iv) .$$

Moreover, for all continuous and compactly supported φ: $\mathbb{R} \to \mathbb{R}$,

$$\mu(\varphi) = \int_{\mathbb{R}} \varphi(x)\mu(dx) = \lim_{v\downarrow 0} \frac{1}{\pi} \int_{\mathbb{R}} \varphi(x)\Im s_\mu(x+iv)dx .$$

In particular, for two continuity points $a < b$ of μ,

$$\mu([a,b]) = \lim_{v\downarrow 0} \frac{1}{\pi} \int_a^b \Im s_\mu(x+iv)dx .$$

The next proposition characterises functions that are Stieltjes transforms of bounded measures on $\mathbb{R}$.

Proposition 2.6 *Assume that the following conditions hold for a complex valued function $g(z)$:*

(i) *g is holomorphic on $\mathbb{C}^+$;*
(ii) *$g(z) \in \mathbb{C}^+$ for all $z \in \mathbb{C}^+$;*

(iii) $\limsup_{v\to\infty} |ivg(iv)| < \infty$.

Then g is the Stieltjes transform of a bounded measure on $\mathbb{R}$.

Similar to the characterisation of the weak convergence of finite measures by the convergence of their Fourier transforms, Stieltjes transform characterises the vague convergence of finite measures. This a key tool for the study of the ESD's of random matrices.

Theorem 2.7 *A sequence $\{\mu_n\}$ of probability measures on $\mathbb{R}$ converges vaguely to some positive measure μ (possibly defective) if and only if their Stieltjes transforms $\{s_{\mu_n}\}$ converges to s_μ on $\mathbb{C}^+$.*

In order to get the weak convergence of $\{\mu_n\}$, one checks the vague convergence of the sequence using this theorem and then to ensure that the limiting measure μ is a probability measure, i.e. to check $\mu(1) = 1$ through Proposition 2.5 or by some direct observation.

The Stieltjes transform and the RMT are closely related each other. Indeed, the Stieltjes transform of the ESD $F^{\mathbf{A}}$ of a $n \times n$ Hermitian matrix $\mathbf{A}$ is by definition

$$s_{\mathbf{A}}(z) = \int \frac{1}{x-z} F^{\mathbf{A}}(dx) = \frac{1}{n} \operatorname{tr}(\mathbf{A} - z\mathbf{I})^{-1} , \tag{2.3}$$

which is the resolvent of the matrix $\mathbf{A}$ (up to the factor $1/n$). Using a formula for the trace of an inverse matrix, see Bai and Silverstein (2010, Theorem A.4), we have

$$s_n(z) = \frac{1}{n} \sum_{k=1}^{n} \frac{1}{a_{kk} - z - \boldsymbol{\alpha}_k^*(\mathbf{A}_k - z\mathbf{I})^{-1}\boldsymbol{\alpha}_k} , \tag{2.4}$$

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 proved to be equal to $g(z, s_n(z)) + o(1)$ for some function g, then a LSD F exists and its Stieltjes transform is the solution to the equation

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

Its applications will be discussed in more detail later in the chapter.

2.3 Marčenko-Pastur distributions

The *Marčenko-Pastur distribution F_{y,σ^2} (M-P law) with index y and scale parameter σ* has the density function

$$p_{y,\sigma^2}(x) = \begin{cases} \frac{1}{2\pi xy\sigma^2} \sqrt{(b-x)(x-a)}, & \text{if } a \le x \le b, \\ 0, & \text{otherwise}, \end{cases} \tag{2.5}$$

with an additional point mass of value $1 - 1/y$ at the origin if $y > 1$, where $a = \sigma^2(1 - \sqrt{y})^2$ and $b = \sigma^2(1 + \sqrt{y})^2$. Here, the constant y is the dimension to sample size ratio index and σ^2 the scale parameter. The distribution has mean σ^2 and variance $y\sigma^4$. The support interval has a length of $b - a = 4\sigma^2\sqrt{y}$.

If $\sigma^2 = 1$, the distribution is said to be a standard M-P distribution (then we simplify the notations to F_y and p_y for the distribution and its density function). Three standard

M-P density functions for $y \in \{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}\}$ are displayed on Figure 2.1. In particular, the density function behaves as $\sqrt{x-a}$ and $\sqrt{b-x}$ at the boundaries a and b, respectively.

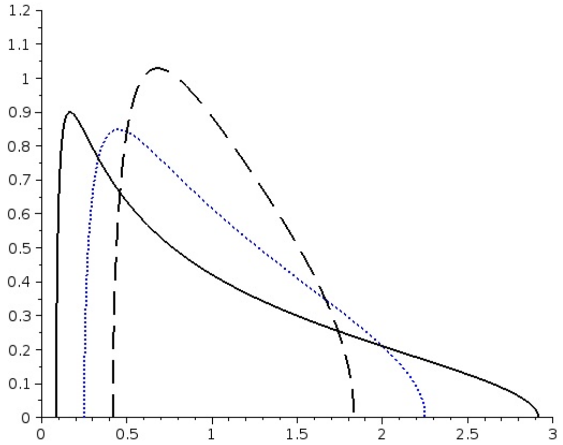


Figure 2.1 Density plots of the Marčenko-Pastur distributions with indexes $y = 1/8$ (dashed line), 1/4 (dotted line) and 1/2 (solid line).

Example 2.8 For the special case of $y = 1$, the density function is

$$p_1(x) = \begin{cases} \frac{1}{2\pi x}\sqrt{x(4-x)}, & \text{if } 0 < x \le 4, \\ 0, & \text{otherwise.} \end{cases} \tag{2.6}$$

In particular, the density is unbounded at the origin.

It is easy to see that when the index y tends to zero, the M-P law F_y shrinks to the Dirac mass δ_1. More intriguing is the following fact (that can be easily checked though): if X_y follows the M-P distribution F_y, then as $y \to 0$, the sequence $\frac{1}{2\sqrt{y}}(X_y - 1)$ weakly converges to Wigner's semi-circle law with density function $\pi^{-1}\sqrt{1-x^2}$ for $|x| \le 1$.

2.3.1 The M-P law for independent vectors without cross-correlations

Notice first that regarding limiting spectral distributions discussed in this chapter, one may ignore the rank-1 matrix $\overline{\mathbf{x}\mathbf{x}}^*$ in the definition of the sample covariance matrix and define the sample covariance matrix to be

$$\mathbf{S}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^*. \tag{2.7}$$

Indeed, by Weilandt-Hoffman inequality, the eigenvalues of the two forms of sample covariance matrix are interlaced each other so that they have a same LSD (when it exists).

As a notational ease, it is also convenient to summarise the n sample vectors into a $p \times n$ random data matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ so that $\mathbf{S}_n = \frac{1}{n}\mathbf{X}\mathbf{X}^*$.

Marčenko and Pastur (1967) first finds the LSD of the large sample covariance matrix $\mathbf{S}_n$. Their result has been extended in various directions afterwards.

Theorem 2.9 *Suppose that the entries $\{x_{ij}\}$ of the data matrix $\mathbf{X}$ are i.i.d. complex random variables with mean zero and variance σ^2, and $p/n \to y \in (0, \infty)$. Then almost surely, $F^{\mathbf{S}_n}$ weakly converges to the MP law F_{y,σ^2} (2.5).*

This theorem was found as early as in the late sixties (convergence in expectation). However its importance for large-dimensional statistics has been recognised only recently at the beginning of this century. To understand its deep influence on multivariate analysis, we plot in Figure 2.2 sample eigenvalues from i.i.d. Gaussian variables $\{x_{ij}\}$. In other words, we use $n = 320$ i.i.d. random vectors $\{\mathbf{x}_i\}$, each with $p = 40$ i.i.d. standard Gaussian coordinates. The histogram of $p = 40$ sample eigenvalues of $\mathbf{S}_n$ displays a wide dispersion from the unit value 1. According to the classical large-sample asymptotic (assuming $n = 320$ is large enough), the sample covariance matrix $\mathbf{S}_n$ should be close to the population covariance matrix $\mathbf{\Sigma} = \mathbf{I}_p = \mathbb{E}\,\mathbf{x}_i\mathbf{x}_i^*$. As eigenvalues are continuous functions of matrix entries, the sample eigenvalues of $\mathbf{S}_n$ should converge to 1 (unique eigenvalue of $\mathbf{I}_p$). The plot clearly assesses that this convergence is far from the reality. On the same graph is also plotted the Marčenko-Pastur density function p_y with $y = 40/320 = 1/8$. The closeness between this density and the sample histogram is striking.

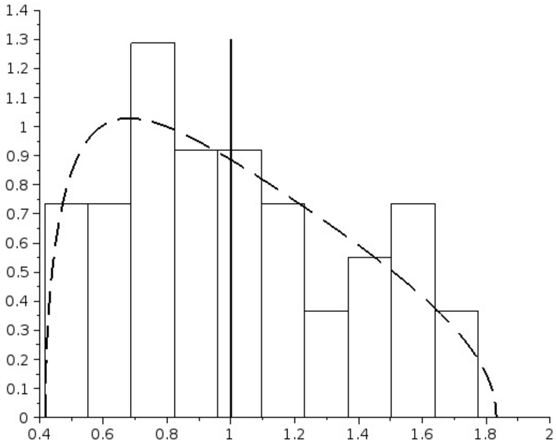


Figure 2.2 Eigenvalues of a sample covariance matrix with standard Gaussian entries, $p = 40$ and $n = 320$. The dashed curve plots the M-P density p_y with $y = 1/8$ and the vertical bar shows the unique population unit eigenvalue.

Since the sample eigenvalues deviate significantly from the population eigenvalues, the sample covariance matrix $\mathbf{S}_n$ is no more a reliable estimator of its population counter-part $\mathbf{\Sigma}$. This observation is indeed the fundamental reason for that classical multivariate methods break down when the data dimension is a bit large compared to the sample size. As an example, consider Hotelling's T^2 statistic which relies on $\mathbf{S}_n^{-1}$. In large-dimensional context (as $p = 40$ and $n = 320$ above), $\mathbf{S}_n^{-1}$ deviates significantly from $\mathbf{\Sigma}^{-1}$. In particular, the wider spread of the sample eigenvalues implies that $\mathbf{S}_n$ may have many small eigenvalues, especially when p/n is close to 1. For example, for $\mathbf{\Sigma} = \sigma^2\mathbf{I}_p$ and $y = 1/8$, the smallest eigenvalue of $\mathbf{S}_n$ is close to $a = (1 - \sqrt{y})^2\sigma^2 = 0.42\sigma^2$ so that the largest eigenvalue of $\mathbf{S}_n^{-1}$ is close to $a^{-1}\sigma^{-2} = 1.55\sigma^{-2}$, a 55% over-spread to the population value σ^{-2}. When the data to sample size increases to $y = 0.9$, the largest eigenvalue of $\mathbf{S}_n^{-1}$ becomes close to $380\sigma^{-2}$! Clearly, $\mathbf{S}_n^{-1}$ is completely unreliable as an estimator of $\mathbf{\Sigma}^{-1}$.

2.3.2 How the Marčenko-Pastur law appears in the limit?

As said in Introduction, most of results in RMT require advanced mathematical tools which are exposed in details elsewhere. Here an explanation why the LSD should be the Marčenko-Pastur distribution is given using Stieltjes transform.

Throughout this book, for a complex number z (or a negative real number), $\sqrt{z}$ denotes its square root with positive imaginary part. Without loss of generality, the scale parameter is fixed to $\sigma^2 = 1$. Let $z = u + iv$ with $v > 0$ and $s(z)$ be the Stieltjes transform of the M-P distribution F_y. From the definition of the density function p_y in (2.5), it is not difficult to find that the Stieltjes transform of the M-P distribution F_y equals

$$s(z) = \frac{(1-y) - z + \sqrt{(z-1-y)^2 - 4y}}{2yz}. \tag{2.8}$$

It is also important to observe that s is a solution in $\mathbb{C}^+$ to the quadratic equation

$$yzs^2 + (z + y - 1)s + 1 = 0. \tag{2.9}$$

Consider the Stieltjes transform $s_n(z)$ of the ESD of $\mathbf{S}_n$ $s_n(z) = p^{-1}\text{tr}(\mathbf{S}_n - z\mathbf{I}_p)^{-1}$. Theorem 2.9 is proved if almost surely $s_n(z) \to s(z)$ for every $z \in \mathbb{C}^+$. Assume that this convergence takes place: what should then be the limit? Since for fixed z, $\{s_n(z)\}$ is bounded, $\mathbb{E}\, s_n(z) \to s(z)$ too.

By Eq.(2.4),

$$s_n(z) = \frac{1}{p}\sum_{k=1}^{p} \frac{1}{\frac{1}{n}\boldsymbol{\alpha}_k'\overline{\boldsymbol{\alpha}}_k - z - \frac{1}{n^2}\boldsymbol{\alpha}_k'\mathbf{X}_k^*(\frac{1}{n}\mathbf{X}_k\mathbf{X}_k^* - z\mathbf{I}_{p-1})^{-1}\mathbf{X}_k\overline{\boldsymbol{\alpha}}_k}\,, \tag{2.10}$$

where $\mathbf{X}_k$ is the matrix obtained from $\mathbf{X}$ with the k-th row removed and $\boldsymbol{\alpha}_k'$ ($n\times 1$) is the k-th row of $\mathbf{X}$. Assume also that all conditions are fulfilled so that the p denominators converge almost surely to their expectations, i.e. the (random) errors caused by this approximation can be controlled to be negligible for large p and n. First,

$$\mathbb{E}\,\frac{1}{n}\boldsymbol{\alpha}_k'\overline{\boldsymbol{\alpha}}_k = \frac{1}{n}\sum_{j=1}^{n}|x_{kj}|^2 = 1.$$

Next,

$$\begin{aligned}
&\mathbb{E}\,\frac{1}{n^2}\boldsymbol{\alpha}_k'\mathbf{X}_k^*(\frac{1}{n}\mathbf{X}_k\mathbf{X}_k^* - z\mathbf{I}_{p-1})^{-1}\mathbf{X}_k\overline{\boldsymbol{\alpha}}_k \\
&= \frac{1}{n^2}\,\mathbb{E}\,\text{tr}\,\mathbf{X}_k^*(\frac{1}{n}\mathbf{X}_k\mathbf{X}_k^* - z\mathbf{I}_{p-1})^{-1}\mathbf{X}_k\overline{\boldsymbol{\alpha}}_k\boldsymbol{\alpha}_k' \\
&= \frac{1}{n^2}\,\text{tr}\left\{\left[\mathbb{E}\,\mathbf{X}_k^*(\frac{1}{n}\mathbf{X}_k\mathbf{X}_k^* - z\mathbf{I}_{p-1})^{-1}\mathbf{X}_k\right]\left[\mathbb{E}\,\overline{\boldsymbol{\alpha}}_k\boldsymbol{\alpha}_k'\right]\right\} \\
&= \frac{1}{n^2}\,\text{tr}\left[\mathbb{E}\,\mathbf{X}_k^*(\frac{1}{n}\mathbf{X}_k\mathbf{X}_k^* - z\mathbf{I}_{p-1})^{-1}\mathbf{X}_k\right] \\
&= \frac{1}{n^2}\,\mathbb{E}\,\text{tr}\left[\mathbf{X}_k^*(\frac{1}{n}\mathbf{X}_k\mathbf{X}_k^* - z\mathbf{I}_{p-1})^{-1}\mathbf{X}_k\right]
\end{aligned}$$

$$= \frac{1}{n^2}\,\mathbb{E}\,\mathrm{tr}\left[(\frac{1}{n}\mathbf{X}_k\mathbf{X}_k^* - z\mathbf{I}_{p-1})^{-1}\mathbf{X}_k\mathbf{X}_k^*\right] .$$

Note that $\frac{1}{n}\mathbf{X}_k\mathbf{X}_k^*$ is a sample covariance matrix close to $\mathbf{S}_n$ (with one vector $\mathbf{x}_k$ removed). Therefore,

$$\begin{aligned}
&\frac{1}{n^2}\,\mathbb{E}\,\mathrm{tr}\left[(\frac{1}{n}\mathbf{X}_k\mathbf{X}_k^* - z\mathbf{I}_{p-1})^{-1}\mathbf{X}_k\mathbf{X}_k^*\right]\\
&\simeq \frac{1}{n^2}\,\mathbb{E}\,\mathrm{tr}\left[(\frac{1}{n}\mathbf{X}\mathbf{X}^* - z\mathbf{I}_p)^{-1}\mathbf{X}\mathbf{X}^*\right]\\
&= \frac{1}{n}\,\mathbb{E}\,\mathrm{tr}\left[(\frac{1}{n}\mathbf{X}\mathbf{X}^* - z\mathbf{I}_p)^{-1}\frac{1}{n}\mathbf{X}\mathbf{X}^*\right]\\
&= \frac{1}{n}\,\mathbb{E}\,\mathrm{tr}\left[I_p + z(\frac{1}{n}\mathbf{X}\mathbf{X}^* - z\mathbf{I}_p)^{-1}\right]\\
&= \frac{p}{n} + z\frac{p}{n}\,\mathbb{E}\,s_n(z) .
\end{aligned}$$

Collecting all these derivations, the expectation of the denominators equal to (up to negligible terms)

$$1 - z - \left\{\frac{p}{n} + z\frac{p}{n}\,\mathbb{E}\,s_n(z)\right\} .$$

On the other hand, the denominators in Eq.(2.10) are bounded above and away from 0 and converge almost surely, the convergence also holds in expectation by the dominating convergence theorem. It is then seen that when $p \to \infty$, $n \to \infty$ and $p/n \to y > 0$, the limit $s(z)$ of $\mathbb{E}\,s_n(z)$ satisfies the equation

$$s(z) = \frac{1}{1 - z - \{y + yzs(z)\}} .$$

This is indeed Eq.(2.9) which characterises the Stieltjes transform of the M-P distribution F_y with index y.

2.3.3 Integrals and moments of the M-P law

It is important to evaluate the integrals of a smooth function f with respect to the standard M-P law in (2.5).

Proposition 2.10 *For the standard Marčenko-Pastur distribution F_y in* (2.5) *with index $y > 0$ and $\sigma^2 = 1$, it holds for any function f analytic on a domain containing the support interval* $[a, b] = [(1 \mp \sqrt{y})^2]$,

$$\int f(x)dF_y(x) = -\frac{1}{4\pi i}\oint_{|z|=1} \frac{f\left(|1 + \sqrt{y}z|^2\right)(1 - z^2)^2}{z^2(1 + \sqrt{y}z)(z + \sqrt{y})}dz. \tag{2.11}$$

This proposition is a corollary of a stronger result, Theorem 2.23, that will be established in Section 2.5.

Example 2.11 Logarithms of eigenvalues are widely used in multivariate analysis. Let $f(x) = \log(x)$ and assume $0 < y < 1$ to avoid null eigenvalues. We will show that

$$\int \log(x)dF_y(x) = -1 + \frac{y-1}{y}\log(1 - y) . \tag{2.12}$$

Indeed, by (2.11),

$$\int \log(x)dF_y(x) = -\frac{1}{4\pi i}\oint_{|z|=1} \frac{\log\left(|1+\sqrt{y}z|^2\right)(1-z^2)^2}{z^2(1+\sqrt{y}z)(z+\sqrt{y})}dz$$

$$= -\frac{1}{4\pi i}\oint_{|z|=1} \frac{\log\left(1+\sqrt{y}z\right)(1-z^2)^2}{z^2(1+\sqrt{y}z)(z+\sqrt{y})}dz$$

$$-\frac{1}{4\pi i}\oint_{|z|=1} \frac{\log\left(1+\sqrt{y}\bar{z}\right)(1-z^2)^2}{z^2(1+\sqrt{y}z)(z+\sqrt{y})}dz.$$

Call the two integrals A and B. For both integrals, the origin is a pole of degree 2, and $-\sqrt{y}$ is a simple pole (recall that $y < 1$). The corresponding residues are respectively

$$\left.\frac{\log\left(1+\sqrt{y}z\right)(1-z^2)^2}{z^2(1+\sqrt{y}z)}\right|_{z=-\sqrt{y}} = \frac{1-y}{y}\log(1-y)\ ,$$

and

$$\left.\frac{\partial}{\partial z}\frac{\log\left(1+\sqrt{y}z\right)(1-z^2)^2}{(1+\sqrt{y}z)(z+\sqrt{y})}\right|_{z=0} = 1\ .$$

Hence by the residue theorem,

$$A = -\frac{1}{2}\left\{\frac{1-y}{y}\log(1-y)+1\right\}.$$

Furthermore,

$$B = -\frac{1}{4\pi i}\oint_{|z|=1} \frac{\log\left(1+\sqrt{y}\bar{z}\right)(1-z^2)^2}{z^2(1+\sqrt{y}z)(z+\sqrt{y})}dz$$

$$= +\frac{1}{4\pi i}\oint_{|\xi|=1} \frac{\log\left(1+\sqrt{y}\xi\right)(1-1/\xi^2)^2}{\frac{1}{\xi^2}(1+\sqrt{y}/\xi)(1/\xi+\sqrt{y})}\cdot -\frac{1}{\xi^2}d\xi \qquad (\text{with } \xi=\bar{z}=1/z)$$

$$= A\ .$$

Hence, the whole integral equals $2A$.

Example 2.12 (mean of the M-P law). We have for all $y > 0$,

$$\int x dF_y(x) = 1\ . \tag{2.13}$$

This can be found in a way similar to Example 2.11. There is however another more direct proof of the result. Indeed almost surely, we have by the weak convergence of the ESD, $p^{-1}\,\mathrm{tr}(\mathbf{S}_n) \to \int x dF_y(x)$. On the other hand,

$$\frac{1}{p}\,\mathrm{tr}(\mathbf{S}_n) = \frac{1}{pn}\sum_{i=1}^{n}\mathrm{tr}[\mathbf{x}_i\mathbf{x}_i^*] = \frac{1}{pn}\sum_{i=1}^{n}\sum_{j=1}^{p}|x_{ij}|^2\ .$$

By the strong law of large numbers, the limit is $\mathbb{E}\,|x_{11}|^2 = 1$.

For a monomial function $f(x) = x^k$ of arbitrary degree k, the residue method of Proposition 2.10 becomes inefficient and a more direct calculation is needed.

Proposition 2.13 *The moments of the standard M-P law ($\sigma^2 = 1$) are*

$$\alpha_k := \int x^k dF_y(x) = \sum_{r=0}^{k-1} \frac{1}{r+1}\binom{k}{r}\binom{k-1}{r}y^r.$$

Proof By definition,

$$\begin{aligned}
\alpha_k &= \frac{1}{2\pi y}\int_a^b x^{k-1}\sqrt{(b-x)(x-a)}dx \\
&= \frac{1}{2\pi y}\int_{-2\sqrt{y}}^{2\sqrt{y}}(1+y+z)^{k-1}\sqrt{4y-z^2}dz \ \text{(with } x = 1+y+z\text{)} \\
&= \frac{1}{2\pi y}\sum_{\ell=0}^{k-1}\binom{k-1}{\ell}(1+y)^{k-1-\ell}\int_{-2\sqrt{y}}^{2\sqrt{y}} z^\ell\sqrt{4y-z^2}dz \\
&= \frac{1}{2\pi y}\sum_{\ell=0}^{[(k-1)/2]}\binom{k-1}{2\ell}(1+y)^{k-1-2\ell}(4y)^{\ell+1}\int_{-1}^{1}u^{2\ell}\sqrt{1-u^2}du, \\
&\qquad \text{(by setting } z = 2\sqrt{y}u\text{)} \\
&= \frac{1}{2\pi y}\sum_{\ell=0}^{[(k-1)/2]}\binom{k-1}{2\ell}(1+y)^{k-1-2\ell}(4y)^{\ell+1}\int_0^1 w^{\ell-1/2}\sqrt{1-w}dw \\
&\qquad \text{(setting } u = \sqrt{w}\text{)} \\
&= \frac{1}{2\pi y}\sum_{\ell=0}^{[(k-1)/2]}\binom{k-1}{2\ell}(1+y)^{k-1-2\ell}(4y)^{\ell+1}\int_0^1 w^{\ell-1/2}\sqrt{1-w}dw \\
&= \sum_{\ell=0}^{[(k-1)/2]}\frac{(k-1)!}{\ell!(\ell+1)!(k-1-2\ell)!}y^\ell(1+y)^{k-1-2\ell} \\
&= \sum_{\ell=0}^{[(k-1)/2]}\sum_{s=0}^{k-1-2\ell}\frac{(k-1)!}{\ell!(\ell+1)!s!(k-1-2\ell-s)!}y^{\ell+s} \\
&= \sum_{\ell=0}^{[(k-1)/2]}\sum_{r=\ell}^{k-1-\ell}\frac{(k-1)!}{\ell!(\ell+1)!(r-\ell)!(k-1-r-\ell)!}y^r \\
&= \frac{1}{k}\sum_{r=0}^{k-1}\binom{k}{r}y^r\sum_{\ell=0}^{\min(r,k-1-r)}\binom{s}{\ell}\binom{k-r}{k-r-\ell-1} \\
&= \frac{1}{k}\sum_{r=0}^{k-1}\binom{k}{r}\binom{k}{r+1}y^r = \sum_{r=0}^{k-1}\frac{1}{r+1}\binom{k}{r}\binom{k-1}{r}y^r.
\end{aligned}$$

□

In particular, $\alpha_1 = 1$, $\alpha_2 = 1 + y$ and the variance of the M-P law equals y.

2.4 Generalised Marčenko-Pastur distributions

In Theorem 2.9, the population covariance matrix has the simplest form $\mathbf{\Sigma} = \sigma^2\mathbf{I}_p$. In order to consider a general population covariance matrix $\mathbf{\Sigma}$, we make the following assumption: the observation vectors $\{\mathbf{y}_k\}_{1\le k\le n}$ can be represented as $\mathbf{y}_k = \mathbf{\Sigma}^{1/2}\mathbf{x}_k$ where the

$\mathbf{x}_k$'s have i.i.d. components as in Theorem 2.9 and $\mathbf{\Sigma}^{1/2}$ is any nonnegative square root of $\mathbf{\Sigma}$. The associated sample covariance matrix is

$$\widetilde{\mathbf{B}}_n = \frac{1}{n}\sum_{k=1}^{n} \mathbf{y}_k\mathbf{y}_k^* = \mathbf{\Sigma}^{1/2}\left(\frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k\mathbf{x}_k^*\right)\mathbf{\Sigma}^{1/2} = \mathbf{\Sigma}^{1/2}\mathbf{S}_n\mathbf{\Sigma}^{1/2} . \tag{2.14}$$

Here $\mathbf{S}_n$ still denotes the sample covariance matrix in (2.7) with i.i.d. components. Note that the eigenvalues of $\widetilde{\mathbf{B}}_n$ are the same as the product $\mathbf{S}_n\mathbf{\Sigma}$.

The following result extends Theorem 2.9 to random matrices of type $\mathbf{B}_n = \mathbf{S}_n\mathbf{T}_n$ for some general nonnegative definite matrix $\mathbf{T}_n$. Such generalisation will be also used for the study of random Fisher matrices where $\mathbf{T}_n$ will be the inverse of an independent sample covariance matrix.

Theorem 2.14 *Let $\mathbf{S}_n$ be the sample covariance matrix defined in* (2.7) *with i.i.d. components and let* $(\mathbf{T}_n)$ *be a sequence of nonnegative definite Hermitian matrices of size $p \times p$. Define $\mathbf{B}_n = \mathbf{S}_n\mathbf{T}_n$ and assume that*

(i) *The entries (x_{jk}) of the data matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ are i.i.d. with mean zero and variance 1;*
(ii) *The data dimension to sample size ratio $p/n \to y > 0$ when $n \to \infty$;*
(iii) *The sequence $(\mathbf{T}_n)$ is either deterministic or independent of $(\mathbf{S}_n)$;*
(iv) *Almost surely, the sequence $(H_n = F^{\mathbf{T}_n})$ of the ESD of $(\mathbf{T}_n)$ weakly converges to a nonrandom probability measure H.*

Then almost surely, $F^{\mathbf{B}_n}$ weakly converges to a nonrandom probability measure $F_{y,H}$. Moreover its Stieltjes transform s is implicitly defined by the equation

$$s(z) = \int \frac{1}{t(1-y-yzs(z))-z}dH(t), \quad z \in \mathbb{C}^+. \tag{2.15}$$

Several important comments are in order. First, it has been proved that the above implicit equation has an unique solution as functions from $\mathbb{C}^+$ onto itself. Second, the solution s has no close-form in general and all information about the LSD $F_{c,H}$ is contained in this equation.

There is however a better way to present the fundamental equation (2.15). Consider for $\mathbf{B}_n$ a *companion matrix*

$$\underline{\mathbf{B}}_n = \frac{1}{n}\mathbf{X}^*\mathbf{T}_n\mathbf{X},$$

which is of size $n \times n$. Both matrices share the same non-null eigenvalues so that their ESD satisfy

$$nF^{\underline{\mathbf{B}}_n} - pF^{\mathbf{B}_n} = (n-p)\delta_0 .$$

Therefore when $p/n \to y > 0$, $F^{\mathbf{B}_n}$ has a limit $F_{c,H}$ if and only if $F^{\underline{\mathbf{B}}_n}$ has a limit $\underline{F}_{c,H}$. In this case, the limits satisfies

$$\underline{F}_{c,H} - yF_{c,H} = (1-y)\delta_0 ,$$

and their respective Stieltjes transforms $\underline{s}$ and s are linked each other by the relation

$$\underline{s}(z) = -\frac{1-y}{z} + ys(z) .$$

Substituting $\underline{s}$ for s in (2.15) yields

$$\underline{s} = -\left(z - y\int \frac{t}{1+t\underline{s}}dH(t)\right)^{-1} .$$

Solving in z leads to

$$z = -\frac{1}{\underline{s}} + y\int \frac{t}{1+t\underline{s}}dH(t) , \tag{2.16}$$

which indeed defines the inverse function of $\underline{s}$.

Although the fundamental equations (2.15) and (2.16) are equivalent each other, we call (2.15) *Marčenko-Pastur equation* and (2.15) *Silverstein equation* for historical reasons. In particular, the inverse map given by Silverstein's equation will be of primary importance for the characterisation of the support of the LSD. Moreover, many inference methods for the limiting spectral distribution H of the population are based on Silverstein equation.

Notice that in most discussions so far on the Stieltjes transform s_μ of a probability measure μ on the real line (such as s for the LSD $F_{y,H}$), the complex variable z is restricted to the upper complex plane $\mathbb{C}^+$. However, such Stieltjes transform is in fact defined on the whole open set $\mathbb{C} \setminus \Gamma_\mu$ where it is analytic, see Proposition 2.4. The restriction to $\mathbb{C}^+$ is mainly for mathematical convenience in that s_μ is a one-to-one map on $\mathbb{C}^+$. This is however not a limitation since properties of s_μ established on $\mathbb{C}^+$ are easily extended to the whole domain $\mathbb{C} \setminus \Gamma_\mu$ by continuity. As an example, both Marčenko-Pastur equation and Silverstein's equation are valid for the whole complex plane excluding the support set Γ of the LSD.

Furthermore, the LSD $F_{y,H}$ and its companion $\underline{F}_{y,H}$ will be called *generalised Marčenko-Pastur distributions* with index (y, H). In the case where $\mathbf{T}_n = \mathbf{\Sigma}$, the LSD H of $\mathbf{\Sigma}$ is called the *population spectral distribution*, or simply PSD. For instance, a *discrete PSD H* with finite support $\{a_1, \ldots, a_k\} \subset \mathbb{R}_+$ is of form

$$H = \sum_{j=1}^{k} t_j \delta_{a_j}, \tag{2.17}$$

where $t_j > 0$ and $t_1 + \cdots + t_k = 1$. This means that the population covariance matrix $\mathbf{\Sigma}$ has approximately eigenvalues $(a_j)_{1\le j\le k}$ of multiplicity $\{[pt_j]\}$, respectively.

Remark 2.15 The standard M-P distribution is easily recovered from the Marčenko-Pastur equations. In this case, $T_n = \mathbf{\Sigma} = \mathbf{I}_p$ so that the PSD $H = \delta_1$ and Eq. (2.15) becomes

$$s(z) = \frac{1}{1 - y - z - yzs(z)} ,$$

which characterises the standard M-P distribution. This is also the unique situation where s possesses a close form and by inversion formula, a density function can be obtained for the corresponding LSD.

Except this simplest case, very few is known about the LSD $F_{y,H}$. An exception is a one-to-one correspondence between the families of their moments given in §2.4.1. An algorithm is also proposed later to compute numerically the density function of the LSD $F_{y,H}$.

2.4.1 Moments and support of a generalised M-P distribution

Lemma 2.16 *The moments $\alpha_j = \int x^j dF_{y,H}(x)$, $j \geq 1$ of the LSD $F_{y,H}$ are linked to the moments $\beta_j = \int t^j dH(t)$ of the PSD H by*

$$\alpha_j = y^{-1} \sum y^{i_1+i_2+\cdots+i_j} (\beta_1)^{i_1} (\beta_2)^{i_2} \cdots (\beta_j)^{i_j} \phi^{(j)}_{i_1,i_2,\cdots,i_j} \tag{2.18}$$

where the sum runs over the following partitions of j:

$$(i_1, \ldots, i_j) \ : \ j = i_1 + 2i_2 + \cdots + ji_j, \quad i_\ell \in \mathbb{N},$$

and $\phi^{(j)}_{i_1,i_2,\cdots,i_j}$ is the multinomial coefficient

$$\phi^{(j)}_{i_1,i_2,\cdots,i_j} = \frac{j!}{i_1! i_2! \cdots i_j! (j + 1 - (i_1 + i_2 + \cdots + i_j))!}. \tag{2.19}$$

This lemma can be proved using the fundamental equation (2.15). As an example, for the first three moments, we have

$$\alpha_1 = \beta_1, \quad \alpha_2 = \beta_2 + y\beta_1^2, \quad \alpha_3 = \beta_3 + 3y\beta_1\beta_2 + y^2\beta_1^3.$$

In particular for the standard M-P law, $H = \delta_{\{1\}}$ so that $\beta_j \equiv 1$ for all $j \geq 0$. Therefore, $\alpha_1 = 1$, $\alpha_2 = 1 + y$ and $\alpha_3 = 1 + 3y + y^2$ as discussed in Section 2.3.3.

In order to derive the support of the LSD $F_{y,H}$, it is sufficient to examine the support of the companion distribution $\underline{F}_{y,H}$. Recall that its Stieltjes transform $\underline{s}(z)$ can be extended to all $z \notin \Gamma_{\underline{F}_{y,H}}$. In particular, for real x outside the support $\Gamma_{\underline{F}_{y,H}}$, $\underline{s}(x)$ is differential and increasing so that we can define an functional inverse $\underline{s}^{-1}$. Moreover, the form of this inverse is already given in Eq. (2.16). It is however more convenient to consider the functional inverse ψ of the function $\alpha : \ x \mapsto -1/\underline{s}(x)$. By (2.16), this inverse function is

$$\psi(\alpha) = \psi_{y,H}(\alpha) = \alpha + y\alpha \int \frac{t}{\alpha - t} dH(t) \,. \tag{2.20}$$

It can be checked that this inverse is indeed well-defined for all $\alpha \notin \Gamma_H$.

Proposition 2.17 *If $\lambda \notin \Gamma_{\underline{F}_{c,H}}$, then $\underline{s}(\lambda) \neq 0$ and $\alpha = -1/\underline{s}(\lambda)$ satisfies*

(i) *$\alpha \notin \Gamma_H$ and $\alpha \neq 0$ (so that $\psi(\alpha)$ is well-defined);*
(ii) *$\psi'(\alpha) > 0$.*

Conversely, if α satisfies (i)-(ii), *then $\lambda = \psi(\alpha) \notin \Gamma_{\underline{F}_{c,H}}$.*

Therefore, Proposition 2.17, establishes the relationship between the supports of the PSD H and of the companion LSD $\underline{F}_{c,H}$. It is then possible to determine the support of $\underline{F}_{c,H}$ by looking at intervals where $\psi' > 0$.

Example 2.18 Consider the LSD $F_{y,H}$ with indexes $y = 0.3$ and H the uniform distribution on the set $\{1, 4, 10\}$. Figure 2.3 displays the corresponding ψ function. The function is strictly increasing on the following intervals: $(-\infty, 0)$, $(0, 0.63)$, $(1.40, 2.57)$ and $(13.19, \infty)$. According to Proposition 2.17, we find that

$$\Gamma^c_{\underline{F}_{y,H}} \cap \mathbb{R}^* = (0, \ 0.32) \cup (1.37, \ 1.67) \cup (18.00, \ \infty).$$

Hence, taking into account that 0 belongs to the support of $\underline{F}_{y,H}$, we have

$$\Gamma_{\underline{F}_{y,H}} = \{0\} \cup [0.32,\ 1.37] \cup [1.67,\ 18.00].$$

Therefore, the support of the LSD $F_{y,H}$ is $[0.32,\ 1.37] \cup [1.67,\ 18.00]$.

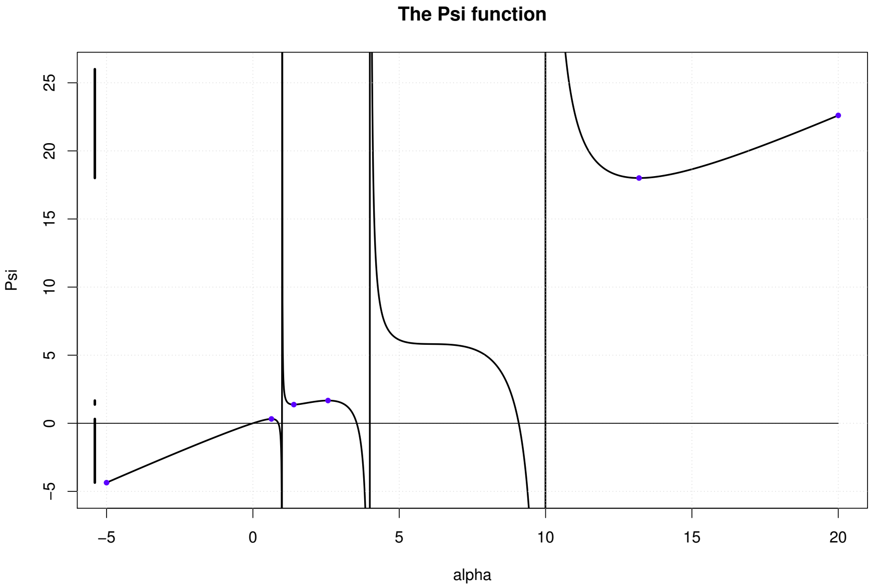


Figure 2.3 The function $\psi_{0.3,H}$ where H is the uniform distribution on $\{1, 4, 10\}$. The dots show the zeros of the derivative and the empty intervals on the broken vertical line on the left are the support of $F_{0.3,H}$.

2.4.2 Numerical computation of a generalised M-P density function

Recall that the Stieltjes transform s of the LSD $F_{y,H}$ is linked to the companion Stieltjes transform $\underline{s}$ via the relationship

$$s = \frac{1}{y}\underline{s} + \frac{1-y}{yz}\ .$$

Let $f_{y,H}$ denote the density function of $F_{y,H}$. By the inversion formula, we have for all $x > 0$,

$$f_{y,H}(x) = \frac{1}{\pi} \lim_{\varepsilon \to 0_+} \Im s(x + i\varepsilon) = \frac{1}{y\pi} \lim_{\varepsilon \to 0_+} \Im \underline{s}(x + i\varepsilon).$$

Numerical approximation of $f_{y,H}(x)$ can then be obtained via the Stieltjes transform $\underline{s}(x + i\varepsilon)$ with very small positive imaginary part, e.g. $\varepsilon = 10^{-6}$.

It remains to approximate the Stieltjes transform and this will be done using the fundamental Silverstein equation (2.16). Rewriting the equation into the form

$$\underline{s} := A(\underline{s}) = \frac{1}{-z + y\int \frac{t}{1+t\underline{s}(z)} dH(t)}\ . \tag{2.21}$$

For a given $z \in \mathbb{C}^+$, $\underline{s}$ is then a fixed point of the map A. Moreover, according to the general random matrix theory, such fixed point exists and is unique on $\mathbb{C}^+$. Therefore, $\underline{s}$ can be found by simply iterating the map A till convergence with an arbitrary point $\underline{s}_0 \in \mathbb{C}^+$. This is referred as the *fixed-point* algorithm for the numerical computation of the Stieltjes transform of a generalised Marčenko-Pastur distributin.

Example 2.19 Consider the LSD $F_{y,H}$ defined in Example 2.18. The computed density function using the above fixed-point algorithm is shown on Figure 2.4. One recovers perfectly the two intervals [0.32,1.37] and [1.67,18.00] of the support. Loosely speaking, the first interval is due to the unit population eigenvalue while the later is due to a mixture effect from the other two population eigenvalues 4 and 10.

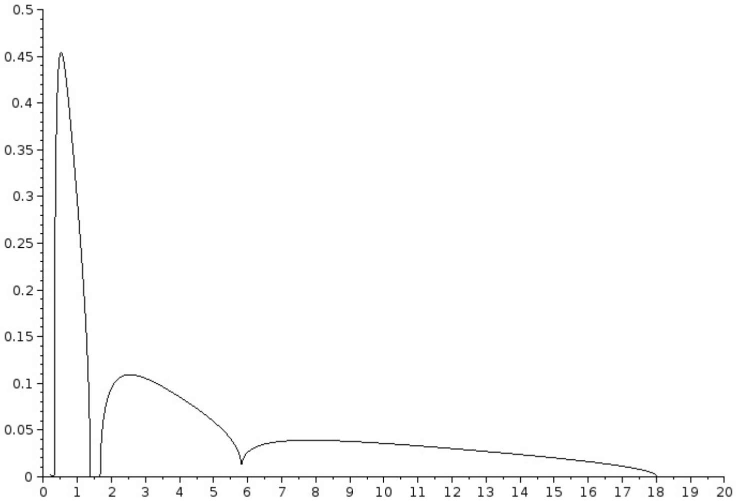


Figure 2.4 The density function for the LSD $F_{0.3,H}$ of Example 2.18 where H is the uniform distribution on $\{1, 4, 10\}$. The support has two intervals: [0.32,1.37] and [1.67,18.00].

Example 2.20 Consider a continuous PSD H defined as the LSD of the Toeplitz matrix $\mathbf{\Sigma} = (2^{-|i-j|})_{1 \le i,j \le p}$ $(p \to \infty)$ and $y = 0.2$. The approximate LSD density $f_{\frac{1}{2},H}$ is given on Figure 2.5. The support of this LSD is a positive compact interval.

2.4.3 Nonparametric estimation of a generalised M-P density function

In a statistical context, the dimension p and the sample size n are both known so that the ratio y can be approximated by p/n. However, the PSD H is unknown and the previous fixed-point algorithm cannot be used to approximate the LSD density $f_{y,H}$. One might first think of an estimation method of H and then compute the LSD density.

Here we present a method using kernel estimation. Indeed, the sample eigenvalues $\lambda_1, \ldots, \lambda_p$ are directly available from the sample covariance matrix $\mathbf{B}_n$. A natural kernel

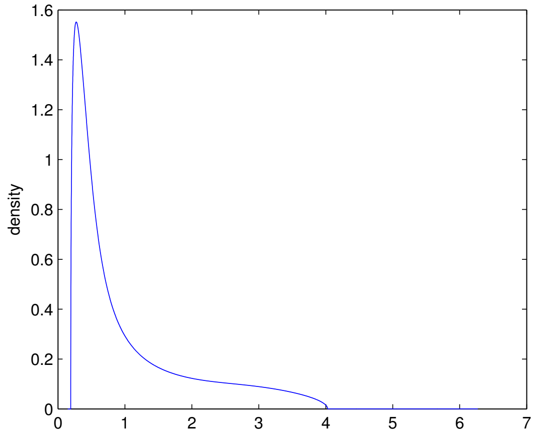


Figure 2.5 Limiting spectral density $f_{\frac{1}{2},H}$ where H is the LSD of a Toeplitz matrix $\mathbf{\Sigma} = (2^{-|i-j|})_{1\le i,j\le p}$ $(p \to \infty)$.

estimator of the LSD density $f_{y,H}$ is therefore

$$\hat{f}_{y,H}(x) = \frac{1}{ph}\sum_{j=1}^{p} K\left(\frac{x-\lambda_j}{h}\right), \tag{2.22}$$

where K is a bounded and nonnegative kernel function satisfying

$$\int K(x)dx = 1, \qquad \int |K'(x)|dx < \infty. \tag{2.23}$$

The estimator $\hat{f}_{y,H}$ is expected to have good asymptotic properties although a rigorous proof of the fact is not straightforward due to the fact that the sample eigenvalues $\{\lambda_j\}$ are dependent.

Theorem 2.21 *In addition to the assumptions in Theorem 2.14, assume that*

(i) *as $n \to \infty$, the window size $h = h(n) \to 0$ satisfying $\lim nh^{5/2} = \infty$;*
(ii) *$\mathbb{E}X_{11}^{16} < \infty$;*
(iii) *the sequence $\{\mathbf{T}_n\}$ is bounded in spectral norm; and*
(iv) *the LSD $F_{y,H}$ has a compact support $[u_1, u_2]$ with $u_1 > 0$.*

Then

$$\hat{f}_{y,H}(x) \to f_{y,H}(x) \quad \textit{in probability and uniformly in } x \in [u_1, u_2].$$

Example 2.22 Let $p = 500, n = 1000$ and we simulate the data with $\mathbf{T}_n = (0.4^{|i-j|})_{1\le i,j\le p}$ and x_{ij}'s are i.i.d $\mathcal{N}(0,1)$-distributed. Figure 2.6 plots a kernel estimate $\hat{f}_{y,H}$ of the LSD density function.

2.5 LSD for random Fisher matrices

For testing the equality between the variances of two Gaussian populations, a Fisher statistic is used which has the form S_1^2/S_2^2 where the S_i^2's are estimators of the unknown vari-

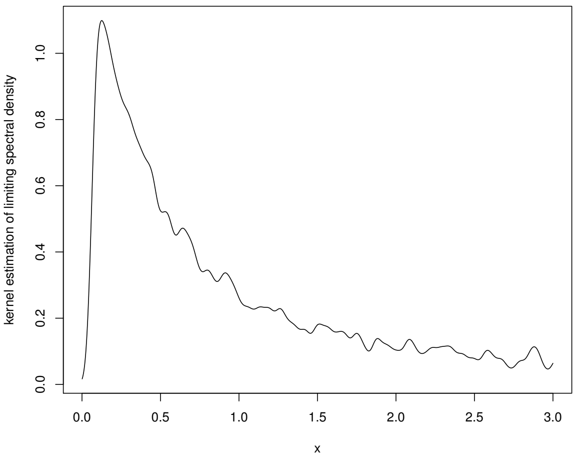


Figure 2.6 Kernel estimation of LSD density $f_{y,H}$ with $p = 100$ and $n = 1000$.

ances in the two populations. The analogue in the multivariate setting is as follows. Consider two independent samples $\{\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}\}$ and $\{\mathbf{y}_1, \ldots, \mathbf{y}_{n_2}\}$, both from a p-dimensional population with i.i.d. components and finite second moment as in Theorem 2.9. Write the respective sample covariance matrices

$$\mathbf{S}_1 = \frac{1}{n_1} \sum_{k=1}^{n_1} \mathbf{x}_k \mathbf{x}_k^*,$$

and

$$\mathbf{S}_2 = \frac{1}{n_2} \sum_{k=1}^{n_2} \mathbf{y}_k \mathbf{y}_k^*.$$

The random matrix

$$\mathbf{F}_n = \mathbf{S}_1 \mathbf{S}_2^{-1}, \tag{2.24}$$

is called a *Fisher matrix* where $n = (n_1, n_2)$ denote the sample sizes. . Since the inverse $\mathbf{S}_2^{-1}$ is used, it is necessary to impose the condition $p \le n_2$ to ensure the invertibility.

In this section, we will derive the LSD of the Fisher matrix $\mathbf{F}_n$.

2.5.1 The Fisher LSD and its integrals

Let $s > 0$ and $0 < t < 1$. The Fisher LSD $F_{s,t}$ is the distribution with the density function

$$p_{s,t}(x) = \frac{1-t}{2\pi x(s+tx)} \sqrt{(b-x)(x-a)}\,, \quad a \le x \le b, \tag{2.25}$$

with

$$a = a(s,t) = \frac{(1-h)^2}{(1-t)^2}\,, \quad b = b(s,t) = \frac{(1+h)^2}{(1-t)^2}\,, \quad h = h(s,t) = (s+t-st)^{1/2}\,. \tag{2.26}$$

Moreover, when $s > 1$, $F_{s,t}$ has a mass at the origin of value $1 - 1/s$ while the total mass of the continuous component above equals $1/s$. Figure 2.7 displays the density functions

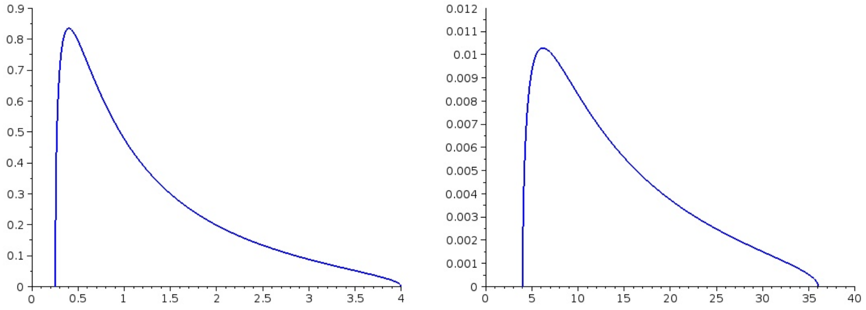


Figure 2.7 Density function of the Fisher LSD distribution $F_{s,t}$. Left: $F_{\frac{1}{5},\frac{1}{5}}$ with support $[\frac{1}{4}, 4]$. Right: $F_{7,\frac{1}{2}}$ has a continuous component on [4,36] and a mass of value $\frac{6}{7}$ at the origin.

of two Fisher LSD, $F_{\frac{1}{5},\frac{1}{5}}$ and $F_{7,\frac{1}{2}}$. In the later case, the distribution has a point mass of value $\frac{6}{7}$ at the origin.

The Fisher LSD share many similarities with the standard Marčenko-Pastur distributions. Indeed, as easily seen from the definition, the standard M-P distribution F_y is the degenerate Fisher LSD $F_{y,0}$ with parameters $(s,t) = (y,0)$. With this connection and by continuity, many distributional calculations done for a Fisher LSD $F_{s,t}$ can be transferred to the M-P distribution by setting $(s,t) = (y,0)$. Notice also that when $t \to 1_-$, $a(s,t) \to \frac{1}{2}(1-s)^2$ while $b(s,t) \to \infty$. The support of the distribution becomes unbounded.

Theorem 2.23 *With the notations given in Theorem 2.28, consider an analytic function f on a domain containing the support interval $[a,b]$. The following formula of contour integral is valid:*

$$\int_a^b f(x)dF_{s,t}(x) = -\frac{h^2(1-t)}{4\pi i}\oint_{|z|=1} \frac{f\left(\frac{|1+hz|^2}{(1-t)^2}\right)(1-z^2)^2 dz}{z(1+hz)(z+h)(tz+h)(t+hz)}.$$

Proof Following (2.25),

$$I = \int_a^b f(x)dF_{s,t}(x) = \int_a^b f(x)\frac{1-t}{2\pi x(s+xt)}\sqrt{(x-a)(b-x)}dx.$$

With the change of variable

$$x = \frac{1+h^2+2h\cos(\theta)}{(1-t)^2},$$

for $\theta \in (0,\pi)$, it holds

$$\sqrt{(x-a)(b-x)} = \frac{2h\sin(\theta)}{(1-t)^2}, \quad dx = \frac{-2h\sin(\theta)}{(1-t)^2}d\theta.$$

Therefore,

$$\begin{aligned} I &= \frac{2h^2(1-t)}{\pi}\int_0^{\pi} \frac{f\left(\frac{(1+h^2+2h\cos(\theta))}{(1-t)^2}\right)\sin^2(\theta)d\theta}{(1+h^2+2h\cos(\theta))(s(1-t)^2+t(1+h^2+2h\cos(\theta)))} \\ &= \frac{h^2(1-t)}{\pi}\int_0^{2\pi} \frac{f\left(\frac{(1+h^2+2h\cos(\theta))}{(1-t)^2}\right)\sin^2(\theta)d\theta}{(1+h^2+2h\cos(\theta))(s(1-t)^2+t(1+h^2+2h\cos(\theta)))}. \end{aligned}$$

Furthermore let $z = e^{i\theta}$, we have

$$1 + h^2 + 2h\cos(\theta) = |1 + hz|^2, \quad \sin(\theta) = \frac{z - z^{-1}}{2i}, \quad d\theta = \frac{dz}{iz}.$$

Consequently,

$$I = -\frac{h^2(1-t)}{4\pi i} \oint_{|z|=1} \frac{f\left(\frac{|1+hz|^2}{(1-t)^2}\right)(1-z^2)^2 dz}{z^3|1+hz|^2(s(1-t)^2 + t|1+hz|^2)}.$$

Finally on the contour it holds $|1 + hz|^2 = (1 + hz)(1 + hz^{-1})$. Substituting it into the above equation leads to the desired result. □

Example 2.24 By taking $(s, t) = (y, 0)$ in Theorem 2.23 we obtain the formula for the Marčenko-Pastur distribution F_y given in Proposition 2.10.

Example 2.25 The first two moments of $F_{s,t}$ are

$$\int x dF_{s,t}(x) = \frac{1}{1-t}, \quad \int x^2 dF_{s,t}(x) = \frac{h^2 + 1 - t}{(1-t)^3}. \tag{2.27}$$

In particular, the variance equals to $h^2/(1-t)^3$. The values are calculated from the integral formula (2.27) with $f(x) = x$ and $f(x) = x^2$, respectively. Notice that by definition, $h > t$ always. For the calculation with x^2, we have

$$\begin{aligned}
&\int_a^b x^2 dF_{s,t}(x) \\
&= -\frac{h^2(1-t)}{4\pi i} \oint_{|z|=1} \frac{\frac{|1+hz|^4}{(1-t)^4}(1-z^2)^2 dz}{z(1+hz)(z+h)(tz+h)(t+hz)} \\
&= -\frac{h^2(1-t)}{4\pi i} \oint_{|z|=1} \frac{(1+hz)(z+h)(1-z^2)^2}{(1-t)^4 z^3 (tz+h)(t+hz)} dz \\
&= -\frac{h^2(1-t)}{4\pi i(1-t)^4 th} \oint_{|z|=1} \frac{(1+hz)(z+h)(1-z^2)^2}{z^3(z+h/t)(z+t/h)} dz \\
&= -\frac{h}{2(1-t)^3 t} \left\{ \left.\frac{(1+hz)(z+h)(1-z^2)^2}{z^3(z+h/t)}\right|_{z=-t/h} + \frac{1}{2}\frac{\partial^2}{\partial z^2} \left.\frac{(1+hz)(z+h)(1-z^2)^2}{(z+h/t)(z+t/h)}\right|_{z=0} \right\} \\
&= \frac{-h}{2t(1-t)^3} \left\{ \frac{(1-t)(h^2-t)(t^2-h)}{ht^2} - 2h - (1 + h^2 - t - h^2/t)\frac{t^2+h^2}{ht} \right\} \\
&= \frac{h^2 + 1 - t}{(1-t)^3}\ .
\end{aligned}$$

Finally, the value of the mean can also be guessed as follows. It should be the limit of $\mathbb{E}[p^{-1} \operatorname{tr} \mathbf{F}_n]$ that is equal to

$$\mathbb{E}[p^{-1} \operatorname{tr} \mathbf{F}_n] = p^{-1} \operatorname{tr} \left\{ \mathbb{E}[\mathbf{S}_1] \cdot \mathbb{E}[\mathbf{S}_2^{-1}] \right\} = \mathbb{E} \left\{ p^{-1} \operatorname{tr} \mathbf{S}_2^{-1} \right\}.$$

As the LSD of $\mathbf{S}_2$ is the M-P law F_t, the limit equals $\int x^{-1} dF_t(x) = (1-t)^{-1}$.

The lemma below gives two more involved moment calculations which will be used later in the book.

Lemma 2.26 *Let*

$$c = c(s,t) = \frac{h}{\sqrt{t}}, \quad d = d(t) = \sqrt{t}\,, \tag{2.28}$$

such that $|1+hz|^2 + st^{-1}(1-t)^2 = |c+dz|^2$ *for all* z. *We have*

$$\int_a^b \frac{x}{x+s/t}dF_{s,t}(x) = \frac{t}{s+t}\,, \tag{2.29}$$

and

$$-\int_a^b \log\left\{(x+s/t)(1-t)^2\right\}dF_{s,t}(x) \tag{2.30}$$

$$= \frac{1-t}{t}\log(c) - \frac{s+t}{st}\log(c-dt/h) + \begin{cases} \frac{(1-s)}{s}\log(c-dh), & 0<s<1,\\ 0, & s=1,\\ -\frac{(1-s)}{s}\log(c-d/h), & s>1. \end{cases} \tag{2.31}$$

Proof For $f(x) = x/(x+s/t)$ and $|z|=1$,

$$\begin{aligned} f\left(\frac{|1+hz|^2}{(1-t)^2}\right) &= \frac{|1+hz|^2/(1-t)^2}{|1+hz|^2/(1-t)^2 + s/t} = \frac{|1+hz|^2}{|c+dz|^2}\\ &= \frac{(1+hz)(h+z)}{(c+dz)(d+cz)} = t\frac{(1+hz)(h+z)}{(h+tz)(t+hz)}. \end{aligned}$$

By (2.27) and noticing that $h>t$,

$$\begin{aligned} \int_0^\infty \frac{x}{x+s/t}dF^{s,t}(x) &= -\frac{h^2t(1-t)}{4\pi i}\oint_{|z|=1}\frac{(1-z^2)^2}{z(tz+h)^2(t+hz)^2}dz\\ &= \frac{t}{s+t}\,. \end{aligned}$$

Next for $f(x) = -\log\left\{(x+s/t)(1-t)^2\right\}$, by (2.27),

$$\begin{aligned} &-\int_a^b \log((x+s/t)(1-t)^2)dF_{s,t}(x)\\ &= \frac{h^2(1-t)}{4\pi i}\oint_{|z|=1}\frac{\log(|c+dz|^2)(1-z^2)^2dz}{z(z+h)(1+hz)(t+hz)(tz+h)} =: A+B, \end{aligned}$$

where

$$\begin{aligned} A &= \frac{h^2(1-t)}{4\pi i}\oint_{|z|=1}\frac{\log(c+dz)(1-z^2)^2dz}{z(z+h)(1+hz)(t+hz)(tz+h)}\,,\\ B &= \frac{h^2(1-t)}{4\pi i}\oint_{|z|=1}\frac{\log(c+d\bar{z})(1-z^2)^2dz}{z(z+h)(1+hz)(t+hz)(tz+h)}\,. \end{aligned}$$

By the variable change $w = \bar{z} = 1/z$ in B, it can be easily proved that $B = A$. First assume $0<s<1$ so that $\sqrt{s} \le h < 1$. Then

$$\begin{aligned} 2A &= \frac{h^2(1-t)}{2\pi i}\oint_{|z|=1}\frac{\log(c+dz)(1-z^2)^2dz}{z(z+h)(1+hz)(t+hz)(tz+h)}\\ &= \frac{(1-t)}{2\pi i\cdot t}\oint_{|z|=1}\frac{\log(c+dz)(1-z^2)^2dz}{z(z+h)(z+1/h)(z+t/h)(z+h/t)} \end{aligned}$$

$$= \frac{1-t}{t}\left\{ \left.\frac{\log(c+dz)(1-z^2)^2}{(z+h)(z+1/h)(z+t/h)(z+h/t)}\right|_{z=0} + \left.\frac{\log(c+dz)(1-z^2)^2}{z(z+1/h)(z+t/h)(z+h/t)}\right|_{z=-h} \right.$$

$$\left. + \left.\frac{\log(c+dz)(1-z^2)^2}{z(z+h)(z+1/h)(z+h/t)}\right|_{z=-t/h} \right\}$$

$$= \frac{1-t}{t}\left\{ \log(c) + \frac{t(1-s)}{s(1-t)}\log(c-dh) - \frac{s+t}{st}\log(c-dt/h) \right\}$$

$$= \frac{1-t}{t}\log(c) + \frac{1-s}{s}\log(c-dh) - \frac{s+t}{st}\log(c-dt/h)\ .$$

When $s > 1$, $1 < h \le \sqrt{s}$, the pole $z = -h$ is replaced by $z = -1/h$ and the corresponding residue by

$$\left.\frac{\log(c+dz)(1-z^2)^2}{z(z+h)(z+t/h)(z+h/t)}\right|_{z=-1/h} = -\frac{t(1-s)}{s(1-t)}\log(c-d/h),$$

so that we obtain in this case

$$2A = \frac{1-t}{t}\log(c) - \frac{1-s}{s}\log(c-d/h) - \frac{s+t}{st}\log(c-dt/h)\ .$$

Finally, the result for $s = 1$ is obtained by continuity. □

2.5.2 Derivation of the LSD of the Fisher matrix $\mathbf{F}_n$

The LSD of $\mathbf{F}_n$ in (2.24) will be derived under the conditions $p/n_1 \to y_1 > 0$ and $p/n_1 \to y_2 \in (0,1)$. By Theorem 2.9, almost surely $F^{\mathbf{S}_2}$ converges to the Marčenko-Pastur distribution F_{y_2} with index y_2. It follows that $F^{\mathbf{S}_2^{-1}}$ converges to a distribution H which equals to the image of F_{y_2} by the transformation $\lambda \mapsto 1/\lambda$ on $(0,\infty)$. By applying theorem 2.14 with $\mathbf{T} = \mathbf{S}_2^{-1}$, it follows that the sequence $(\mathbf{F}_n)$ has a LSD denoted by μ. Let s be its Stieltjes transform and $\underline{s}$ be the companion Stieltjes transform (of the measure $y_1\mu + (1-y_1)\delta_0$). Then $\underline{s}$ satisfies the Marčenko-Pastur equation

$$z = -\frac{1}{\underline{s}} + y_1\int\frac{t}{1+t\underline{s}}dH(t) = -\frac{1}{\underline{s}} + y_1\int\frac{1}{t+\underline{s}}dF_{y_2}(t).$$

This identity can be rewritten as

$$z + \frac{1}{\underline{s}} = y_1\overline{s_2(-\underline{s})},$$

where $s_2(z)$ denotes the Stieltjes transform of the M-P distribution F_{y_2}. Using its value given in Eq.(2.8) leads to

$$\bar{z} + \frac{1}{\bar{\underline{s}}} = y_1\frac{1-y_2+\bar{\underline{s}}+\sqrt{(1+y_2+\bar{\underline{s}})^2-4y_2}}{-2y_2\bar{\underline{s}}}\ .$$

Take the square and after some arrangements,

$$\bar{z}(y_1+y_2\bar{z})\bar{\underline{s}}^2 + [\bar{z}(y_1+2y_2-y_1y_2)+y_1-y_1^2]\bar{\underline{s}} + y_1+y_2-y_1y_2 = 0\ .$$

By taking the conjugate and solving in $\underline{s}$ leads to, with $h^2 = y_1+y_2-y_1y_2$,

$$\underline{s}(z) = -\frac{z(h^2+y_2)+y_1-y_1^2-y_1\sqrt{(z(1-y_2)-1+y_1)^2-4zh^2}}{2z(y_1+y_2z)}\ .$$

Moreover, the density function of the LSD μ can be found as follows:

$$\begin{aligned} p_{y_1,y_2}(x) &= \frac{1}{\pi}\Im(s(x+i0)) = \frac{1}{y_1\pi}\Im(\underline{s}(x+i0)) \\ &= \frac{1-y_2}{2\pi x(y_1+y_2x)}\sqrt{(b-x)(x-a)}\,, \end{aligned}$$

for $x \in [a, b]$, a finite interval depending on the indexes y_1 and y_2 only. Furthermore, in case of $y_1 > 1$, intuitively $\mathbf{S}_1$ has $p - n_1$ null eigenvalues while $\mathbf{S}_2$ is of full rank, $\mathbf{F}$ should also has $p - n_1$ null eigenvalues. Consequently, the LSD μ should have a point mass of value $1 - 1/y_1$ at the origin. This can be rigorously proved with the formula

$$G(\{0\}) = -\lim_{z\to 0+0i} zs_G(z)\,,$$

which is valid for any probability measure on $[0,\infty)$. Applying to μ yields

$$\mu(\{0\}) = -\lim_{z\to 0+0i} zs(z) = 1 - \frac{1}{y_1} - \frac{1}{y_1}\lim_{z\to 0+0i} z\underline{s}(z) = 1 - \frac{1}{y_1}\,, \tag{2.32}$$

which proves the conjecture.

Remark 2.27 In case of $y_1 \leq 1$, the above computation is still valid and the LSD μ has no mass at the origin.

These results are summarised in the following theorem.

Theorem 2.28 *Let $\mathbf{S}_k$, $k = 1, 2$ be p-th dimensional sample covariance matrices from two independent samples $\{\mathbf{x}_i\}_{1\leq i\leq n_1}$ and $\{\mathbf{y}_j\}_{1\leq j\leq n_2}$ of sizes n_1 and n_2, respectively, both of the type given in Theorem 2.9 with i.i.d. components of unit variance. Consider the Fisher matrix $\mathbf{F}_n = \mathbf{S}_1\mathbf{S}_2^{-1}$ and let*

$$n_1 \to \infty, \quad n_2 \to \infty, \quad p/n_1 \to y_1 > 0\,, \quad p/n_2 \to y_2 \in (0,1)\,.$$

Then, almost surely, the ESD of $\mathbf{F}_n$ weakly converges to the Fisher LSD F_{y_1,y_2} with parameters (y_1, y_2).

In conclusion, a Fisher LSD is the limiting spectral distribution of a random Fisher matrix.

Notes

For a general introduction to related random matrix theory, we recommend the monographs by Tao (2012), Bai and Silverstein (2010), Anderson et al. (2010) and Pastur and Shcherbina (2011). In particular, Tao (2012) provides an introduction at the graduate level while the other texts are more specialised. Related to the topics developed in this book, Bai (1999) gives a quick review of the major tools and idea involved.

The celebrated Marčenko-Pastur distributions as well as the Marčenko-Pastur equation (2.15) first appeared in Marčenko and Pastur (1967). The Silverstein equation (2.16) establishing the inverse map of the Stieltjes transform of the LSD is due to J. Silverstein and appears first in Silverstein and Combettes (1992). As explained in the chapter, this equation is instrumental for the derivation of many results presented in this book.

Lemma 2.16 can be proved using either the Marčenko-Pastur equation (2.15) or the

Silverstein equation (2.16). For an alternative proof, see Nica and Speicher (2006, page 143).

Proposition 2.17 is established in Silverstein and Choi (1995). More information on the support of the LSD $F_{c,H}$ can be found in this paper. For example, for a finite discrete PSD H with k masses, the support of the LSD $F_{c,H}$ has at most k compact intervals.

Theorem 2.21 is due to Jing et al. (2010).

3

CLT for linear spectral statistics

3.1 Introduction

In Chapter 2, the sample covariance matrices $\mathbf{S}_n$, $\mathbf{B}_n$ and the sample Fisher random matrix $\mathbf{F}_n$ are introduced and their limiting spectral distributions are found under some general conditions. Let $\mathbf{A}_n$ be one of these sample matrices. In one-sample and two-sample multivariate analysis, many statistics are functions of the eigenvalues $\{\lambda_k\}$ of the sample matrix $\mathbf{A}_n$ of form

$$T_n = \frac{1}{p}\sum_{k=1}^{p}\varphi(\lambda_k) = \int \varphi(x) dF^{\mathbf{A}_n}(x) =: F^{\mathbf{A}_n}(\varphi) \,, \tag{3.1}$$

for some specific function φ. Such statistic is called a *linear spectral statistic* of the sample matrix $\mathbf{A}_n$.

Example 3.1 The *generalised variance* discussed in Chapter 4, see Eq.(4.1) is

$$T_n = \frac{1}{p}\log|\mathbf{S}_n| = \frac{1}{p}\sum_{k=1}^{p}\log(\lambda_k).$$

So T_n is a simple linear spectral statistic of the sample covariance matrix $\mathbf{S}_n$ with $\varphi(x) = \log(x)$.

Example 3.2 To test the hypothesis $H_0 : \mathbf{\Sigma} = \mathbf{I}_p$ that the population covariance matrix is equal to a given matrix, the log-likelihood ratio statistic (assuming a Gaussian population) is

$$LRT_1 = \operatorname{tr}\mathbf{S}_n - \log|\mathbf{S}_n| - p = \sum_{k=1}^{p}[\lambda_k - \log(\lambda_k) - 1] \,.$$

This test is detailed in § 5.6.1. The test statistic is thus p-times a linear spectral statistic of the sample covariance matrix with $\varphi(x) = x - \log(x) - 1$.

Example 3.3 For the two-sample test of the hypothesis $H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ that two populations have a common covariance matrix, the log-likelihood ratio statistic (assuming Gaussian populations) is

$$LRT_2 = -\log|\mathbf{I}_p + \alpha_n\mathbf{F}_n| = -\sum_{k=1}^{p}[1 + \alpha_n\log(\lambda_k)] \,,$$

where α_n is some constant (depending on the sample sizes). This test is presented in

§ 5.6.4. The test statistic is thus p-times a linear spectral statistic of the random Fisher matrix with $\varphi(x) = -\log(1 + \alpha_n x)$.

When the dimension and the sample size tend to infinity in a proportional way, the sample matrix $\mathbf{A}_n$ has a LSD, say F, as discussed in Chapter 2. Since this LSD has bounded support, we have then for any continuous function φ (as the ones given in the examples above), $F^{\mathbf{A}_n}(\varphi) \rightarrow F(\varphi)$ almost surely. How to characterise the fluctuation of $F^{\mathbf{A}_n}(\varphi)$ around its limit $F(\varphi)$? The central limit theorems in this chapter address this issue.

3.2 CLT for linear spectral statistics of a sample covariance matrix

In this section, we consider the simplest sample covariance matrix $\mathbf{S}_n$ defined in (2.7) and satisfying the conditions in Theorem 2.9, i.e. the $p \times n$ data matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) = (x_{ij})$ is then made with np i.i.d standardised entries with $\mathbb{E}\, x_{ij} = 0$, $\mathbb{E}\, |x_{ij}|^2 = 1$ and if the variables are complex-valued, $\mathbb{E}\, x_{ij}^2 = 0$. The LSD of $\mathbf{S}_n$ is known to be the standard Marčenko-Pastur law F_y with index $y = \lim p/n$. In particular, almost surely, $F^{\mathbf{S}_n}(\varphi) \rightarrow F_y(\varphi)$ for continuous function φ.

In order to go a step further, a "natural" way would be to consider the difference $F^{\mathbf{S}_n}(\varphi) - F_y(\varphi)$, that is the fluctuation of $F^{\mathbf{S}_n}(\varphi)$ around its limit. However, from the random matrix theory, it is known that for smooth function, typically the fluctuation $F^{\mathbf{S}_n}(\varphi)$ *around its mean* is of order $1/p$, i.e. $p\left\{F^{\mathbf{S}_n}(\varphi) - \mathbb{E}\, F^{\mathbf{S}_n}(\varphi)\right\}$ converges to a Gaussian distribution. From the decomposition,

$$p\left\{F^{\mathbf{S}_n}(\varphi) - F_y(\varphi)\right\} = p\left\{F^{\mathbf{S}_n}(\varphi) - \mathbb{E}\, F^{\mathbf{S}_n}(\varphi)\right\} + p\left\{\mathbb{E}\, F^{\mathbf{S}_n}(\varphi) - F_y(\varphi)\right\},$$

we see that the fluctuation around the limit $F_y(\varphi)$ depends on the order of the bias $\mathbb{E}\, F^{\mathbf{S}_n}(\varphi) - F_y(\varphi)$. Furthermore, this bias is typically a function of $y_n - y = p/n - y$, the difference between the dimension-to-sample ratio y_n and its limit y. Since $y_n - y$ can have an arbitrary order, e.g. $y_n - y \propto p^{-\alpha}$ for arbitrary $\alpha > 0$, because of the multiplication of the bias by p, the last term in the above decomposition can blow up to infinity (for small α), tend to a constant, or converges to zero (for large α). Therefore, it is not possible to characterise the fluctuation around the limit $F_y(\varphi)$ without specifying further conditions on the difference $y_n - y$.

On the other hand, it is difficult to determine accurately the value of $\mathbb{E}\, F^{\mathbf{S}_n}(\varphi)$. A successful solution to this problem is to consider the fluctuation $F^{\mathbf{S}_n}(\varphi) - F_{y_n}(\varphi)$, that is around $F_{y_n}(\varphi)$, a *finite-horizon proxy* for the limit $F_y(\varphi)$ obtained by substituting the current dimension-to-sample ratio y_n for its limit value y.

In all the following, we use an indicator κ set to 2 when $\{x_{ij}\}$ are real and to 1 when they are complex. Define

$$\beta = E|x_{ij}|^4 - 1 - \kappa, \quad h = \sqrt{y}. \tag{3.2}$$

The coefficient β is indeed the fourth-cumulant of the entries $\{x_{ij}\}$. In particular, if the variables are Gaussian, $\beta = 0$. Recall that by Eq.(2.16), the Stieltjes transform $\underline{s}$ of the companion distribution $\underline{F}_y = (1 - y)\delta_0 + yF_y$ satisfies the Marčenko-Pastur equation

$$z = -\frac{1}{\underline{s}} + \frac{y}{1 + \underline{s}}, \quad z \in \mathbb{C}^+. \tag{3.3}$$

Theorem 3.4 *Assume that the variables $\{x_{ij}\}$ of the data matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ are independent and identically distributed satisfying $Ex_{ij} = 0$, $E|x_{ij}|^2 = 1$, $E|x_{ij}|^4 = \beta + 1 + \kappa < \infty$ and in case of complex variables, $Ex_{ij}^2 = 0$. Assume moreover,*

$$p \to \infty, \quad n \to \infty, \quad p/n \to y > 0\,.$$

Let $f_1, \cdots f_k$ be functions analytic on an open region containing the support of F_y. The random vector $\{X_n(f_1), \cdots X_n(f_k)\}$ where

$$X_n(f) = p\left\{F^{\mathbf{S}_n}(f) - F_{y_n}(f)\right\}$$

converges weakly to a Gaussian vector $(X_{f_1}, \cdots X_{f_k})$ with mean function and covariance function:

$$\mathbb{E}[X_f] = (\kappa - 1)I_1(f) + \beta I_2(f)\,, \tag{3.4}$$

$$cov(X_f, X_g) = \kappa J_1(f,g) + \beta J_2(f,g)\,, \tag{3.5}$$

where

$$I_1(f) = -\frac{1}{2\pi i}\oint \frac{y\left\{\underline{s}/(1+\underline{s})\right\}^3(z)f(z)}{\left[1 - y\left\{\underline{s}/(1+\underline{s})\right\}^2\right]^2}dz\,,$$

$$I_2(f) = -\frac{1}{2\pi i}\oint \frac{y\left\{\underline{s}/(1+\underline{s})\right\}^3(z)f(z)}{1 - y\left\{\underline{s}/(1+\underline{s})\right\}^2}dz\,,$$

$$J_1(f,g) = -\frac{1}{4\pi^2}\oint\oint \frac{f(z_1)g(z_2)}{(\underline{m}(z_1) - \underline{m}(z_2))^2}\underline{m}'(z_1)\underline{m}'(z_2)dz_1dz_2\,,$$

$$J_2(f,g) = \frac{-y}{4\pi^2}\oint f(z_1)\frac{\partial}{\partial z_1}\left\{\frac{\underline{s}}{1+\underline{s}}(z_1)\right\}dz_1 \cdot \oint g(z_2)\frac{\partial}{\partial z_2}\left\{\frac{\underline{s}}{1+\underline{s}}(z_2)\right\}dz_2\,,$$

where the integrals are along contours (non overlapping in J_1) enclosing the support of F_y.

A noticeable feature in this CLT is that the asymptotic mean $\mathbb{E}[X_f]$ is in general non null and its value depends on the forth cumulant of the distributions of the entries. While the LSD, namely the Marčenko-Pastur distribution depends only on the value of the second moment of the entries, the CLT for linear spectral statistics depends on the first four moments of the distribution.

Remark 3.5 In Theorem 3.4 and for complex-valued functions $\{f_j\}$, $(X_{f_1}, \ldots, X_{f_k})$ is said to follow a Gaussian distribution in the sense that its real and imaginary parts have a joint $(2k)$-dimensional real Gaussian distribution. This differs from the standard definition of a complex-valued Gaussian vector. Moreover, the covariance function is defined in this theorem to be

$$\mathrm{cov}(X_f, X_g) = \mathbb{E}\{X_f - \mathbb{E}X_f\}\{X_g - \mathbb{E}X_g\}\,.$$

Furthermore, the variance of X_f is computed as $\mathrm{cov}(X_f, X_{\bar{f}})$ where if $f(z) = u(z) + iv(z)$, $\bar{f}(z) = u(z) - iv(z)$. Note that with this definition, f is analytic if and only if $\bar{f}$ does.

However, concrete applications of Theorem 3.4 are not easy since the limiting parameters are given through those integrals on contours that are only vaguely defined. The following Proposition convert all these integrals to integrals along the unit circle. These formula are much easier to use for concrete applications, see for example Proposition 3.8 below.

Proposition 3.6 *The limiting parameters in Theorem 3.4 can be expressed as follows:*

$$I_1(f) = \lim_{r\downarrow 1} I_1(f,r)\ , \tag{3.6}$$

$$I_2(f) = \frac{1}{2\pi i}\oint_{|\xi|=1} f(|1+h\xi|^2)\frac{1}{\xi^3}d\xi\ , \tag{3.7}$$

$$J_1(f,g) = \lim_{r\downarrow 1} J_1(f,g,r)\ , \tag{3.8}$$

$$J_2(f,g) = -\frac{1}{4\pi^2}\oint_{|\xi_1|=1}\frac{f(|1+h\xi_1|^2)}{\xi_1^2}d\xi_1\oint_{|\xi_2|=1}\frac{g(|1+h\xi_2|^2)}{\xi_2^2}d\xi_2\ , \tag{3.9}$$

with

$$I_1(f,r) = \frac{1}{2\pi i}\oint_{|\xi|=1} f(|1+h\xi|^2)[\frac{\xi}{\xi^2-r^{-2}}-\frac{1}{\xi}]d\xi\ ,$$

$$J_1(f,g,r) = -\frac{1}{4\pi^2}\oint_{|\xi_1|=1}\oint_{|\xi_2|=1}\frac{f(|1+h\xi_1|^2)g(|1+h\xi_2|^2)}{(\xi_1-r\xi_2)^2}d\xi_1 d\xi_2\ .$$

Proof We start with the simplest formula $I_2(f)$ to explain the main argument and indeed, the other formulas are obtained similarly. The idea is to introduce the change of variable $z = 1 + hr\xi + hr^{-1}\overline{\xi} + h^2$ with $r > 1$ but close to 1 and $|\xi| = 1$ (recall $h = \sqrt{y}$). It can be readily checked that when ξ runs anticlockwise the unit circle, z will run a contour C that encloses closely the support interval $[a,b] = [(1 \pm h)^2]$. Moreover, by Eq. (3.3), we have on C

$$\underline{s} = -\frac{1}{1+hr\xi},\quad \text{and } dz = h(r - r^{-1}\xi^{-2})d\xi\ .$$

Applying this variable change to the formula of $I_2(f)$ given in Theorem 3.4, we have

$$\begin{aligned} I_2(f) &= \lim_{r\downarrow 1}\frac{1}{2\pi i}\oint_{|\xi|=1} f(z)\frac{1}{\xi^3}\frac{r\xi^2-r^{-1}}{r(r^2\xi^2-1)}d\xi \\ &= \frac{1}{2\pi i}\oint_{|\xi|=1} f(|1+h\xi|^2)\frac{1}{\xi^3}d\xi\ . \end{aligned}$$

This proves the formula (3.7). For (3.6), we have similarly

$$\begin{aligned} I_1(f) &= \lim_{r\downarrow 1}\frac{1}{2\pi i}\oint_{|\xi|=1} f(z)\frac{1}{\xi^3}\frac{r\xi^2-r^{-1}}{r(r^2\xi^2-1)}\frac{1}{1-r^{-2}\xi^{-2}}d\xi \\ &= \lim_{r\downarrow 1}\frac{1}{2\pi i}\oint_{|\xi|=1} f(|1+h\xi|^2)\frac{1}{\xi(\xi^2-r^{-2})} \\ &= \lim_{r\downarrow 1} I_1(f,r)\ . \end{aligned}$$

Formula (3.9) for $J_2(f,g)$ is calculated in a same fashion by observing that we have

$$\frac{\partial}{\partial z}\left\{\frac{\underline{s}}{1+\underline{s}}(z)\right\}dz = \frac{\partial}{\partial \xi}\left\{\frac{\underline{s}}{1+\underline{s}}(\xi)\right\}d\xi = \frac{\partial}{\partial \xi}\left\{\frac{1}{-hr\xi}\right\}d\xi = \frac{1}{hr\xi^2}d\xi\ .$$

Finally for (3.8), we use two non-overlapping contours defined by $z_j = 1+hr_j\xi_j+hr_j^{-1}\overline{\xi}_j+$

h^2, $j = 1, 2$ where $r_2 > r_1 > 1$. By observing that

$$\underline{s}'(z_j)dz_j = \left(\frac{\partial}{\partial \xi_j}\underline{s}\right)d\xi_j = \frac{hr_j}{(1 + hr_j\xi_j)^2}d\xi_j \ ,$$

we find

$$\begin{aligned} J_1(f,g) = & \lim_{\substack{r_2 > r_1 > 1 \\ r_2 \downarrow 1}} -\frac{1}{4\pi^2}\oint_{|\xi_1|=1}\oint_{|\xi_2|=1}\frac{f(z_1)g(z_2)}{\left\{\underline{s}(z_1) - \underline{s}(z_2)\right\}^2} \\ & \cdot\frac{hr_1}{(1 + hr_1\xi_1)^2}\cdot\frac{hr_2}{(1 + hr_2\xi_2)^2}d\xi_1 d\xi_2 \\ = & \lim_{\substack{r_2 > r_1 > 1, \\ r_2 \downarrow 1}} -\frac{1}{4\pi^2}\oint_{|\xi_1|=1}\oint_{|\xi_2|=1}\frac{f(z_1)g(z_2)}{\{r_1\xi_1 - r_2\xi_2\}^2}d\xi_1 d\xi_2 \\ = & \lim_{r\downarrow 1} -\frac{1}{4\pi^2}\oint_{|\xi_1|=1}\oint_{|\xi_2|=1}\frac{f(|1 + h\xi_1|^2)g(|1 + h\xi_2|^2)}{\{\xi_1 - r\xi_2\}^2}d\xi_1 d\xi_2 \ . \end{aligned}$$

The proof is complete. □

Remark 3.7 There is another useful formula for the integral $I_1(f)$ in the limiting mean $\mathbb{E}X_f$:

$$I_1(f) = \frac{f((1 - \sqrt{y})^2) + f((1 + \sqrt{y})^2)}{4} - \frac{1}{2\pi}\int_{(1-\sqrt{y})^2}^{(1+\sqrt{y})^2}\frac{f(x)}{\sqrt{4y - (x - 1 - y)^2}}dx. \qquad (3.10)$$

3.2.1 A detailed application of the CLT

Proposition 3.8 *Consider two linear spectral statistics*

$$\sum_{i=1}^{p}\log(\lambda_i), \quad \sum_{i=1}^{p}\lambda_i$$

where $\{\lambda_i\}$ *are the eigenvalues of the sample covariance matrix* $\mathbf{S}_n$. *Then under the assumptions of Theorem 3.4,*

$$\begin{pmatrix} \sum_{i=1}^{p}\log\lambda_i - pF_{y_n}(\log x) \\ \sum_{i=1}^{p}\lambda_i - pF_{y_n}(x) \end{pmatrix} \Longrightarrow \mathcal{N}(\mu_1, V_1),$$

with

$$\mu_1 = \begin{pmatrix} \frac{\kappa-1}{2}\log(1 - y) - \frac{1}{2}\beta y \\ 0 \end{pmatrix},$$

$$V_1 = \begin{pmatrix} -\kappa\log(1 - y) + \beta y & (\beta + \kappa)y \\ (\beta + \kappa)y & (\beta + \kappa)y \end{pmatrix}, \quad \textit{and}$$

$$F_{y_n}(x) = 1, \quad F_{y_n}(\log x) = \frac{y_n - 1}{y_n}\log 1 - y_n - 1 \ .$$

Proof Let for $x > 0$, $f(x) = \log x$ and $g(x) = x$. Applying Theorem 3.4 to the pair (f, g) gives

$$\begin{pmatrix} \sum_{i=1}^{p}\log\lambda_i - pF_{y_n}(\log x) \\ \sum_{i=1}^{p}\lambda_i - pF_{y_n}(x) \end{pmatrix} \Longrightarrow \mathcal{N}\left(\begin{pmatrix} EX_f \\ EX_g \end{pmatrix}, \begin{pmatrix} \mathrm{cov}(X_f, X_f) & \mathrm{cov}(X_f, X_g) \\ \mathrm{cov}(X_g, X_f) & \mathrm{cov}(X_g, X_g) \end{pmatrix}\right).$$

Firstly, the values of centring parameters $F_{y_n}(\log x)$ and $F_{y_n}(x)$ are calculated in Examples 2.11 and 2.12. It remains to evaluate the limiting parameters using Proposition 3.6. They are found from the following calculations where h is denoted as $\sqrt{y}$:

$$I_1(f,r) = \frac{1}{2}\log\left(1 - h^2/r^2\right), \tag{3.11}$$

$$I_1(g,r) = 0, \tag{3.12}$$

$$I_2(f) = -\frac{1}{2}h^2, \tag{3.13}$$

$$I_2(g) = 0, \tag{3.14}$$

$$J_1(f,g,r) = \frac{h^2}{r^2}, \tag{3.15}$$

$$J_1(f,f,r) = -\frac{1}{r}\log(1 - h^2/r), \tag{3.16}$$

$$J_1(g,g,r) = \frac{h^2}{r^2}, \tag{3.17}$$

$$J_2(f,g) = h^2, \tag{3.18}$$

$$J_2(f,f) = h^2, \tag{3.19}$$

$$J_2(g,g) = h^2. \tag{3.20}$$

Proof of (3.11): We have

$$\begin{aligned}
I_1(f,r) &= \frac{1}{2\pi i}\oint_{|\xi|=1} f(|1+h\xi|^2)\Big[\frac{\xi}{\xi^2 - r^{-2}} - \frac{1}{\xi}\Big]d\xi \\
&= \frac{1}{2\pi i}\oint_{|\xi|=1} \log(|1+h\xi|^2)\Big[\frac{\xi}{\xi^2 - r^{-2}} - \frac{1}{\xi}\Big]d\xi \\
&= \frac{1}{2\pi i}\oint_{|\xi|=1} (\frac{1}{2}\log((1+h\xi)^2) + \frac{1}{2}\log((1+h\xi^{-1})^2)\Big[\frac{\xi}{\xi^2 - r^{-2}} - \frac{1}{\xi}\Big]d\xi \\
&= \frac{1}{2\pi i}\Big[\oint_{|\xi|=1} \log(1+h\xi)\frac{\xi}{\xi^2 - r^{-2}}d\xi - \oint_{|\xi|=1} \log(1+h\xi)\frac{1}{\xi}d\xi \\
&\quad + \oint_{|\xi|=1} \log(1+h\xi^{-1})\frac{\xi}{\xi^2 - r^{-2}}d\xi - \oint_{|\xi|=1} \log(1+h\xi^{-1})\frac{1}{\xi}d\xi\Big].
\end{aligned}$$

For the first integral, note that as $r > 1$, the poles are $\pm\frac{1}{r}$ and we have by the residue theorem,

$$\begin{aligned}
&\frac{1}{2\pi i}\oint_{|\xi|=1} \log(1+h\xi)\frac{\xi}{\xi^2 - r^{-2}}d\xi \\
&= \frac{\log(1+h\xi)\cdot\xi}{\xi - r^{-1}}\bigg|_{\xi=-r^{-1}} + \frac{\log(1+h\xi)\cdot\xi}{\xi + r^{-1}}\bigg|_{\xi=r^{-1}} \\
&= \frac{1}{2}\log(1 - \frac{h^2}{r^2})\,.
\end{aligned}$$

For the second integral,

$$\frac{1}{2\pi i}\oint_{|\xi|=1} \log(1+h\xi)\frac{1}{\xi}d\xi = \log(1+h\xi)\big|_{\xi=0} = 0\,.$$

The third integral is

$$\begin{aligned}
&\frac{1}{2\pi i}\oint_{|\xi|=1} \log(1+h\xi^{-1})\frac{\xi}{\xi^2 - r^{-2}}d\xi \\
&= -\frac{1}{2\pi i}\oint_{|z|=1} \log(1+hz)\frac{z^{-1}}{z^{-2} - r^{-2}}\cdot\frac{-1}{z^2}dz \\
&= \frac{1}{2\pi i}\oint_{|z|=1} \frac{\log(1+hz)r^2}{z(z+r)(z-r)}dz = \frac{\log(1+hz)r^2}{(z+r)(z-r)}\bigg|_{z=0} = 0\,,
\end{aligned}$$

where the first equality results from the change of variable $z = \frac{1}{\xi}$, and the third equality holds because $r > 1$, so the only pole is $z = 0$. Finally, the fourth integral equals

$$\begin{aligned}&\frac{1}{2\pi i}\oint_{|\xi|=1}\log(1+h\xi^{-1})\frac{1}{\xi}d\xi = -\frac{1}{2\pi i}\oint_{|z|=1}\log(1+hz)\frac{-z}{z^2}dz\\ &= \log(1+hz)\big|_{z=0} = 0\ .\end{aligned}$$

Collecting the four integrals leads to the desired formula for $I_1(f, r)$.

Proof of (3.12): We have

$$\begin{aligned}I_1(g,r) &= \frac{1}{2\pi i}\oint_{|\xi|=1} g(|1+h\xi|^2)\cdot[\frac{\xi}{\xi^2-r^{-2}}-\frac{1}{\xi}]d\xi\\ &= \frac{1}{2\pi i}\oint_{|\xi|=1}|1+h\xi|^2\cdot\Big[\frac{\xi}{\xi^2-r^{-2}}-\frac{1}{\xi}\Big]d\xi\\ &= \frac{1}{2\pi i}\oint_{|\xi|=1}\frac{\xi+h+h\xi^2+h^2\xi}{\xi}\cdot\Big[\frac{\xi}{\xi^2-r^{-2}}-\frac{1}{\xi}\Big]d\xi\\ &= \frac{1}{2\pi i}\oint_{|\xi|=1}\frac{\xi+h+h\xi^2+h^2\xi}{\xi^2-r^{-2}}d\xi - \frac{1}{2\pi i}\oint_{|\xi|=1}\frac{\xi+h+h\xi^2+h^2\xi}{\xi^2}d\xi\ .\end{aligned}$$

These two integrals are calculated as follows:

$$\begin{aligned}&\frac{1}{2\pi i}\oint_{|\xi|=1}\frac{\xi+h+h\xi^2+h^2\xi}{\xi^2-r^{-2}}d\xi\\ &= \frac{\xi+h+h\xi^2+h^2\xi}{\xi-r^{-1}}\Big|_{\xi=-r^{-1}} + \frac{\xi+h+h\xi^2+h^2\xi}{\xi+r^{-1}}\Big|_{\xi=r^{-1}} = 1+h^2\ ;\end{aligned}$$

and

$$\frac{1}{2\pi i}\oint_{|\xi|=1}\frac{\xi+h+h\xi^2+h^2\xi}{\xi^2}d\xi = \frac{\partial}{\partial\xi}(\xi+h+h\xi^2+h^2\xi)\Big|_{\xi=0} = 1+h^2\ .$$

Therefore, $I_1(g, r) = 0$.

Proof of (3.13):

$$\begin{aligned}I_2(f) &= \frac{1}{2\pi i}\oint_{|\xi|=1}\log(|1+h\xi|^2)\frac{1}{\xi^3}d\xi\\ &= \frac{1}{2\pi i}\Bigg[\oint_{|\xi|=1}\frac{\log(1+h\xi)}{\xi^3}d\xi + \oint_{|\xi|=1}\frac{\log(1+h\xi^{-1})}{\xi^3}d\xi\Bigg]\ .\end{aligned}$$

We have

$$\begin{aligned}&\frac{1}{2\pi i}\oint_{|\xi|=1}\frac{\log(1+h\xi)}{\xi^3}d\xi = \frac{1}{2}\frac{\partial^2}{\partial\xi^2}\log(1+h\xi)\Big|_{\xi=0} = -\frac{1}{2}h^2\ ;\\ &\frac{1}{2\pi i}\oint_{|\xi|=1}\frac{\log(1+h\xi^{-1})}{\xi^3}d\xi = -\frac{1}{2\pi i}\oint_{|z|=1}\frac{\log(1+hz)}{\frac{1}{z^3}}\cdot\frac{-1}{z^2}dz = 0\ .\end{aligned}$$

Combining the two leads to $I_2(f) = -\frac{1}{2}h^2$.

Proof of (3.14):

$$I_2(g) = \frac{1}{2\pi i}\oint_{|\xi|=1}\frac{(1+h\overline{\xi})(1+h\xi)}{\xi^3}d\xi = \frac{1}{2\pi i}\oint_{|\xi|=1}\frac{\xi+h\xi^2+h+h^2\xi}{\xi^4}d\xi = 0\ .$$

Proof of (3.15):

$$J_1(f,g,r) = \frac{1}{2\pi i}\oint_{|\xi_2|=1} |1+h\xi_2|^2 \cdot \frac{1}{2\pi i}\oint_{|\xi_1|=1} \frac{\log(|1+h\xi_1|^2)}{(\xi_1 - r\xi_2)^2} d\xi_1 d\xi_2 \,.$$

We have,

$$\begin{aligned}
&\frac{1}{2\pi i}\oint_{|\xi_1|=1} \frac{\log(|1+h\xi_1|^2)}{(\xi_1 - r\xi_2)^2} d\xi_1 \\
&= \frac{1}{2\pi i}\oint_{|\xi_1|=1} \frac{\log(1+h\xi_1)}{(\xi_1 - r\xi_2)^2} d\xi_1 + \frac{1}{2\pi i}\oint_{|\xi_1|=1} \frac{\log(1+h\xi_1^{-1})}{(\xi_1 - r\xi_2)^2} d\xi_1 \,.
\end{aligned}$$

The first term

$$\frac{1}{2\pi i}\oint_{|\xi_1|=1} \frac{\log(1+h\xi_1)}{(\xi_1 - r\xi_2)^2} d\xi_1 = 0,$$

because for fixed $|\xi_2| = 1$, $|r\xi_2| = |r| > 1$, so $r\xi_2$ is not a pole. The second term is

$$\begin{aligned}
\frac{1}{2\pi i}\oint_{|\xi_1|=1} \frac{\log(1+h\xi_1^{-1})}{(\xi_1 - r\xi_2)^2} d\xi_1 &= -\frac{1}{2\pi i}\oint_{|z|=1} \frac{\log(1+hz)}{(\frac{1}{z} - r\xi_2)^2} \cdot \frac{-1}{z^2} dz \\
&= \frac{1}{2\pi i} \cdot \frac{1}{(r\xi_2)^2} \oint_{|z|=1} \frac{\log(1+hz)}{(z - \frac{1}{r\xi_2})^2} dz = \frac{1}{(r\xi_2)^2} \cdot \frac{\partial}{\partial z} \log(1+hz)\Big|_{z=\frac{1}{r\xi_2}} \\
&= \frac{h}{r\xi_2(r\xi_2 + h)} \,,
\end{aligned}$$

where the first equality results from the change of variable $z = \frac{1}{\xi_1}$, and the third equality holds because for fixed $|\xi_2| = 1$, $|\frac{1}{r\xi_2}| = \frac{1}{|r|} < 1$, so $\frac{1}{r\xi_2}$ is a pole of second order.

Therefore,

$$\begin{aligned}
&J_1(f,g,r) \\
&= \frac{h}{2\pi i r^2}\oint_{|\xi_2|=1} \frac{(1+h\xi_2)(1+h\overline{\xi_2})}{\xi_2(\xi_2 + \frac{h}{r})} d\xi_2 \\
&= \frac{h}{2\pi i r^2}\oint_{|\xi_2|=1} \frac{\xi_2 + h\xi_2^2 + h + h^2\xi_2}{\xi_2^2(\xi_2 + \frac{h}{r})} d\xi_2 \\
&= \frac{h}{2\pi i r^2}\left[\oint_{|\xi_2|=1} \frac{1+h^2}{\xi_2(\xi_2 + \frac{h}{r})} d\xi_2 + \oint_{|\xi_2|=1} \frac{h}{\xi_2 + \frac{h}{r}} d\xi_2 + \oint_{|\xi_2|=1} \frac{h}{\xi_2^2(\xi_2 + \frac{h}{r})} d\xi_2\right] .
\end{aligned}$$

Finally we find $J_1(f,g,r) = \frac{h^2}{r^2}$ since

$$\begin{aligned}
&\frac{h}{2\pi i r^2}\oint_{|\xi_2|=1} \frac{1+h^2}{\xi_2(\xi_2 + \frac{h}{r})} d\xi_2 = 0\,, \quad \frac{h}{2\pi i r^2}\oint_{|\xi_2|=1} \frac{h}{\xi_2 + \frac{h}{r}} d\xi_2 = \frac{h^2}{r^2}\,, \\
&\frac{h}{2\pi i r^2}\oint_{|\xi_2|=1} \frac{h}{\xi_2^2(\xi_2 + \frac{h}{r})} d\xi_2 = 0\,.
\end{aligned}$$

Proof of (3.16):

$$J_1(f,f,r) = \frac{1}{2\pi i}\oint_{|\xi_2|=1} f(|1+h\xi_2|^2) \cdot \frac{1}{2\pi i}\oint_{|\xi_1|=1} \frac{f(|1+h\xi_1|^2)}{(\xi_1 - r\xi_2)^2} d\xi_1 d\xi_2$$

$$= \frac{1}{2\pi i}\oint_{|\xi_2|=1} f(|1+h\xi_2|^2)\frac{h}{r\xi_2(r\xi_2+h)}d\xi_2$$
$$= \frac{h}{2\pi i r^2}\oint_{|\xi_2|=1} \frac{\log(1+h\xi_2)}{\xi_2(\frac{h}{r}+\xi_2)}d\xi_2 + \frac{h}{2\pi i r^2}\oint_{|\xi_2|=1} \frac{\log(1+h\xi_2^{-1})}{\xi_2(\frac{h}{r}+\xi_2)}d\xi_2 \ .$$

We have

$$\frac{h}{2\pi i r^2}\oint_{|\xi_2|=1} \frac{\log(1+h\xi_2)}{\xi_2(\frac{h}{r}+\xi_2)}d\xi_2$$
$$= \frac{h}{r^2}\left[\frac{\log(1+h\xi_2)}{\frac{h}{r}+\xi_2}\bigg|_{\xi_2=0} + \frac{\log(1+h\xi_2)}{\xi_2}\bigg|_{\xi_2=-\frac{h}{r}}\right]$$
$$= -\frac{1}{r}\log(1-\frac{h^2}{r}) \ ,$$

and

$$\frac{h}{2\pi i r^2}\oint_{|\xi_2|=1} \frac{\log(1+h\xi_2^{-1})}{\xi_2(\frac{h}{r}+\xi_2)}d\xi_2 = \frac{-h}{2\pi i r^2}\oint_{|z|=1} \frac{\log(1+hz)}{\frac{1}{z}(\frac{h}{r}+\frac{1}{z})}\cdot\frac{-1}{z^2}dz$$
$$= \frac{1}{2\pi i r}\oint_{|z|=1} \frac{\log(1+hz)}{z+\frac{r}{h}}dz = 0 \ ,$$

where the first equality results from the change of variable $z = \frac{1}{\xi_2}$, and the third equality holds because $|\frac{r}{h}| > 1$, so $\frac{r}{h}$ is not a pole.
Finally, we find $J_1(f,f,r) = -\frac{1}{r}\log(1-\frac{h^2}{r})$.

Proof of (3.17):

$$J_1(g,g,r) = \frac{1}{2\pi i}\oint_{|\xi_2|=1} |1+h\xi_2|^2 \cdot \frac{1}{2\pi i}\oint_{|\xi_1|=1} \frac{|1+h\xi_1|^2}{(\xi_1-r\xi_2)^2}d\xi_1 d\xi_2 \ .$$

We have

$$\frac{1}{2\pi i}\oint_{|\xi_1|=1} \frac{|1+h\xi_1|^2}{(\xi_1-r\xi_2)^2}d\xi_1 = \frac{1}{2\pi i}\oint_{|\xi_1|=1} \frac{\xi_1+h\xi_1^2+h+h^2\xi_1}{\xi_1(\xi_1-r\xi_2)^2}d\xi_1$$
$$= \frac{1}{2\pi i}\left[\oint_{|\xi_1|=1} \frac{1+h^2}{(\xi_1-r\xi_2)^2}d\xi_1 + \oint_{|\xi_1|=1} \frac{h\xi_1}{(\xi_1-r\xi_2)^2}d\xi_1 \right.$$
$$\left. + \oint_{|\xi_1|=1} \frac{h}{\xi_1(\xi_1-r\xi_2)^2}d\xi_1\right]$$
$$= \frac{h}{r^2\xi_2^2} \ ,$$

since

$$\frac{1}{2\pi i}\oint_{|\xi_1|=1} \frac{1+h^2}{(\xi_1-r\xi_2)^2}d\xi_1 = 0 \ , \quad \frac{1}{2\pi i}\oint_{|\xi_1|=1} \frac{h\xi_1}{(\xi_1-r\xi_2)^2}d\xi_1 = 0 \ ,$$
$$\frac{1}{2\pi i}\oint_{|\xi_1|=1} \frac{h}{\xi_1(\xi_1-r\xi_2)^2}d\xi_1 = \frac{h}{(\xi_1-r\xi_2)^2}\bigg|_{\xi_1=0} = \frac{h}{r^2\xi_2^2} \ .$$

The last equality holds because for fixed $|\xi_2| = 1$, $|r\xi_2| = |r| > 1$, so $r\xi_2$ is not a pole.

Therefore,

$$\begin{aligned}
J_1(g,g,r) &= \frac{h}{2\pi i r^2}\oint_{|\xi_2|=1}\frac{\xi_2+h\xi_2^2+h+h^2\xi_2}{\xi_2^3}d\xi_2\\
&= \frac{h}{2\pi i r^2}\left[\oint_{|\xi_2|=1}\frac{1+h^2}{\xi_2^2}d\xi_2+\oint_{|\xi_2|=1}\frac{h}{\xi_2}d\xi_2+\oint_{|\xi_2|=1}\frac{h}{\xi_2^3}d\xi_2\right],\\
&= \frac{h^2}{r^2}\,.
\end{aligned}$$

Proof of (3.18) (3.19) (3.20): We have

$$\begin{aligned}
&\frac{1}{2\pi i}\oint_{|\xi_1|=1}\frac{f(|1+h\xi_1|^2)}{\xi_1^2}d\xi_1 = \frac{1}{2\pi i}\oint_{|\xi_1|=1}\frac{\log(|1+h\xi_1|^2)}{\xi_1^2}d\xi_1\\
&= \frac{1}{2\pi i}\oint_{|\xi_1|=1}\frac{\log(1+h\xi_1)+\log(1+h\xi_1^{-1})}{\xi_1^2}d\xi_1 \ = h\,,
\end{aligned}$$

since

$$\begin{aligned}
&\frac{1}{2\pi i}\oint_{|\xi_1|=1}\frac{\log(1+h\xi_1)}{\xi_1^2}d\xi_1 = \frac{\partial}{\partial\xi_1}\Big(\log(1+h\xi_1)\Big)\Big|_{\xi_1=0} = h\,,\\
&\frac{1}{2\pi i}\oint_{|\xi_1|=1}\frac{\log(1+h\xi_1^{-1})}{\xi_1^2}d\xi_1 = -\frac{1}{2\pi i}\oint_{|z|=1}\frac{\log(1+hz)}{\frac{1}{z^2}}\cdot(-\frac{1}{z^2}dz)\\
&= \frac{1}{2\pi i}\oint_{|z|=1}\log(1+hz)dz = 0\,.
\end{aligned}$$

Similarly,

$$\frac{1}{2\pi i}\oint_{|\xi_2|=1}\frac{g(|1+h\xi_2|^2)}{\xi_2^2}d\xi_2 = \frac{1}{2\pi i}\oint_{|\xi_2|=1}\frac{\xi_2+h\xi_2^2+h+h^2\xi_2}{\xi_2^3}d\xi_2 = h.$$

Therefore,

$$\begin{aligned}
J_2(f,g) &= \frac{1}{2\pi i}\oint_{|\xi_1|=1}\frac{f(|1+h\xi_1|^2)}{\xi_1^2}d\xi_1\cdot\frac{1}{2\pi i}\oint_{|\xi_2|=1}\frac{g(|1+h\xi_2|^2)}{\xi_2^2}d\xi_2 = h^2\,,\\
J_2(f,f) &= \frac{1}{2\pi i}\oint_{|\xi_1|=1}\frac{f(|1+h\xi_1|^2)}{\xi_1^2}d\xi_1\cdot\frac{1}{2\pi i}\oint_{|\xi_2|=1}\frac{f(|1+h\xi_2|^2)}{\xi_2^2}d\xi_2 = h^2\,,\\
J_2(g,g) &= \frac{1}{2\pi i}\oint_{|\xi_1|=1}\frac{g(|1+h\xi_1|^2)}{\xi_1^2}d\xi_1\cdot\frac{1}{2\pi i}\oint_{|\xi_2|=1}\frac{g(|1+h\xi_2|^2)}{\xi_2^2}d\xi_2 = h^2\,.
\end{aligned}$$

□

3.3 Bai and Silverstein's CLT

The CLT in Theorem 3.4 assumes the simplest LSD, namely the Marčenko-Pastur law F_y, so that the population covariance matrix $\mathbf{\Sigma}$ is asymptotically close to the identity matrix. The following CLT allows a general population covariance as in Theorem 2.14 that leads to generalised Marčenko-Pastur distributions. Therefore, consider the sample covariance matrix $\widetilde{\mathbf{B}}_n = \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{S}_n\mathbf{\Sigma}^{\frac{1}{2}}$ defined in (2.14), or equivalently the random matrix $\mathbf{B}_n = \mathbf{S}_n\mathbf{T}_n$

defined in Theorem 2.14. Under the conditions of this theorem, the ESD of $\mathbf{B}_n$ converges to the generalised Marčenko-Pastur distribution $F_{y,H}$.

Theorem 3.9 *Let $\{x_{ij}\}$ be the variables in the data matrix $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. Assume that the following conditions hold:*

(a) the variables $\{x_{ij}\}$ are i.i.d., $\mathbb{E}x_{ij} = 0$, $\mathbb{E}|x_{ij}|^2 = 1$, $\mathbb{E}|x_{ij}|^4 < \infty$.

(b) $p \wedge n \to \infty$ and $y_n := p/n \to y > 0$.

(c) $\mathbf{T}_n$ is $p \times p$ nonrandom Hermitian nonnegative definite with spectral norm bounded in p, and its ESD $H_n = F^{\mathbf{T}_n}$ converges weakly to a non-random probability distribution H.

Let $f_1, \cdots, f_k$ be functions analytic on an open region containing the interval

$$\left[\liminf_n \lambda_{\min}^{\mathbf{T}_n} I_{(0,1)}(y)(1-\sqrt{y})^2, \limsup_n \lambda_{\max}^{\mathbf{T}_n}(1+\sqrt{y})^2\right]. \tag{3.21}$$

Then the random vector $(X_n(f_1), \cdots, X_n(f_k))$ where

$$X_n(f) = p\left\{F^{\mathbf{B}_n}(f) - F_{y_n,H_n}\right\}$$

converges weakly to a Gausssian vector $(X_{f_1}, \ldots, X_{f_k})$ whose mean and covariance functions are determined as follows.

(i) *If the x_{ij}'s and $\mathbf{T}_n$ are real and $\mathbb{E}(x_{ij}^4) = 3$, the mean function is*

$$\mathbb{E}X_f = -\frac{1}{2\pi i}\oint_C f(z)\frac{y\int \frac{\underline{s}(z)^3 t^2 dH(t)}{(1+t\underline{s}(z))^3}}{\left(1 - y\int \frac{\underline{s}(z)^2 t^2 dH(t)}{(1+t\underline{s}(z))^2}\right)^2}dz, \tag{3.22}$$

and the covariance function is

$$cov(X_f, X_g) = -\frac{1}{2\pi^2}\oint_{C_1}\oint_{C_2}\frac{f(z_1)g(z_2)}{(\underline{s}(z_1) - \underline{s}(z_2))^2}\underline{s}'(z_1)\underline{s}'(z_2)dz_1dz_2, \tag{3.23}$$

where the integrals are along contours (non-overlapping for the covariance function) which are closed and positively oriented and enclosing the support of $F^{y,H}$.

(ii) *If the x_{ij}'s are complex with $\mathbb{E}(x_{ij}^2) = 0$ and $\mathbb{E}(|x_{ij}|^4) = 2$, the means function is identically zero and the covariance function is $1/2$ times the function given above for the real case.*

Compared to the previous CLT in Theorem 3.4, as explained earlier, the new CLT has the advantage that a general population covariance matrix $\mathbf{S}$ is allowed. However, this CLT has a limitation: the entries $\{x_{ij}\}$ are assumed to have a Gaussian-like 4-th moment while in Theorem 3.4, this moment can be arbitrary.

3.4 CLT for linear spectral statistics of random Fisher matrices

Consider the random Fisher matrix $\mathbf{F}_n$ defined in (2.24) and satisfying the conditions of Theorem 2.28. Denote its ESD by $F_{\mathbf{n}} := F^{\mathbf{F}_n}$ where $\mathbf{n} = (n_1, n_2)$ are the sizes of the two sample $(\mathbf{x}_1, \ldots, \mathbf{x}_{n_1})$ and $(\mathbf{y}_1, \ldots, \mathbf{y}_{n_2})$. The two dimension-to-sample ratios are denoted as $y_{n_1} = p/n_1$ and $y_{n_2} = p/n_2$ and they converge to $(y_1, y_2) \in (0, \infty)\times(0, 1)$. By Theorem 2.28, almost surely $F_{\mathbf{n}}$ converges weakly to the Fisher LSD F_{y_1,y_2} defined in Eqs.(2.25)-(2.26). Consequently, for all continuous function φ, almost surely the linear spectral statistic $F_{\mathbf{n}}(\varphi)$ converges to $F_{y_1,y_2}(\varphi)$.

For the same reasons as in §3.2 for the sample covariance matrix, the fluctuation of the linear spectral statistic $F_{\mathbf{n}}(\varphi)$ cannot be studied around the limit $F_{y_1,y_2}(\varphi)$, but around some finite-horizon proxy of the limit, namely the value $F_{y_{n_1},y_{n_2}}(\varphi)$ obtained from the limit by substituting the current dimension-to-sample ratios (y_{n_1}, y_{n_2}) for their limits (y_1, y_2) in the Fisher LSD.

Let $s(z)$ be the Stieltjes transform of the Fisher LSD F_{y_1,y_2} and define a companion Stieltjes transform $\underline{s}(z) = -\frac{1-y_1}{z} + y_1 s(z)$. Let $s_{y_2}(z)$ be the Stieltjes transform of the Marčenko-Pastur law F_{y_2} (LSD of the covariance matrix $\mathbf{S}_2$) and $\underline{s}_{y_2}(z) = -\frac{1-y_2}{z} + y_2 s_{y_2}(z)$ be its companion Stieltjes transform. Finally, define

$$m_0(z) = \underline{s}_{y_2}(-\underline{s}(z)). \tag{3.24}$$

Again the complex-valued and real-valued cases are distinguished using the indicator variable κ: $\kappa = 1$ when all the variables are complex-valued, and $\kappa = 2$ when they are all real-valued.

Theorem 3.10 *Assume that*

(i) *the two samples* $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}) = (x_{ij})$ *and* $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_{n_2}) = (y_{ik})$ *are as in Theorem 2.28 made with i.i.d. entries, respectively;*

(ii) $\mathbb{E}\, x_{ij} = 0$, $\mathbb{E}\, |x_{ij}|^2 = 1$, $E|x_{jk}|^4 = \beta_x + 1 + \kappa + o(1) < \infty$ *and* $\mathbb{E}\, y_{ik} = 0$, $\mathbb{E}\, |y_{ik}|^2 = 1$, $E|y_{jk}|^4 = \beta_y + 1 + \kappa + o(1) < \infty$. *And in case there are complex-valued,* $\mathbb{E}\, x_{ij}^2 = \mathbb{E}\, y_{ik}^2 = 0$.

(iii) *the dimension p and the sample sizes* (n_1, n_2) *tend to infinity such that*

$$y_{n_1} := p/n_1 \to y_1 \in (0, +\infty), \qquad y_{n_2} := p/n_2 \to y_2 \in (0, 1)\ , \tag{3.25}$$

Let $f_1, \cdots, f_k$ *be k analytic functions on an open domain of the complex plane enclosing the interval* $[a, b]$, *which is the support of the continuous component of the Fisher LSD* F_{y_1,y_2}. *Then, as* $\mathbf{n} \to \infty$, *the random vector* $(X_n(f_1), \ldots, X_n(f_k))$ *where*

$$X_n(f) := p\left\{F_{\mathbf{n}}(f) - F_{y_{n_1},y_{n_2}}(f)\right\},$$

converges weakly to a Gaussian vector $(X_{f_1}, \cdots, X_{f_k})$ *with mean function*

$$\begin{aligned}
\mathbb{E}X_{f_i} = {} & \frac{\kappa - 1}{4\pi i} \oint f_i(z)\ \ d\log\left(\frac{(1-y_2)m_0^2(z) + 2m_0(z) + 1 - y_1}{(1-y_2)m_0^2(z) + 2m_0(z) + 1}\right) \\
& + \frac{\kappa - 1}{4\pi i} \oint f_i(z)\ \ d\log\left(1 - y_2 m_0^2(z)(1 + m_0(z))^{-2}\right) \\
& + \frac{\beta_x \cdot y_1}{2\pi i} \oint f_i(z)\,(m_0(z) + 1)^{-3}\, dm_0(z) \\
& + \frac{\beta_y}{4\pi i} \oint f_i(z)\left(1 - y_2 m_0^2(z)(1 + m_0(z))^{-2}\right) d\log\left(1 - y_2 m_0^2(z)(1 + m_0(z))^{-2}\right),
\end{aligned} \tag{3.26}$$

and covariance functions

$$\begin{aligned}
cov(X_{f_i}, X_{f_j}) = {} & -\frac{\kappa}{4\pi^2} \oint\oint \frac{f_i(z_1)f_j(z_2))dm_0(z_1)dm_0(z_2)}{(m_0(z_1) - m_0(z_2))^2} \\
& - \frac{(\beta_x y_1 + \beta_y y_2)}{4\pi^2} \oint\oint \frac{f_i(z_1)f_j(z_2)dm_0(z_1)dm_0(z_2)}{(m_0(z_1) + 1)^2(m_0(z_2) + 1)^2}.
\end{aligned} \tag{3.27}$$

Again, it is worth noticing that the limiting parameters depend on the fourth cumulants of the variables contrary to the Fisher LSD that depends only on their second moments. Next, similarly as in Proposition3.6, it is possible to calculate the limiting mean and covariance functions using contour integrals on the unit circle.

Proposition 3.11 *The limiting mean and covariance functions in Theorem 3.10 can be determined as*

$$\begin{aligned}\mathbb{E}X_{f_i} = \lim_{r\downarrow 1}\frac{\kappa-1}{4\pi i}\oint_{|\xi|=1} f_i\left(\frac{1+h^2+2h\Re(\xi)}{(1-y_2)^2}\right)\left[\frac{1}{\xi-r^{-1}}+\frac{1}{\xi+r^{-1}}-\frac{2}{\xi+\frac{y_2}{h}}\right]\,d\xi\\ +\frac{\beta_x\cdot y_1(1-y_2)^2}{2\pi i\cdot h^2}\oint_{|\xi|=1} f_i\left(\frac{1+h^2+2h\Re(\xi)}{(1-y_2)^2}\right)\frac{1}{(\xi+\frac{y_2}{h})^3}\,d\xi\ ,\\ +\frac{\beta_y\cdot(1-y_2)}{4\pi i}\oint_{|\xi|=1} f_i\left(\frac{1+h^2+2h\Re(\xi)}{(1-y_2)^2}\right)\frac{\xi^2-\frac{y_2}{h^2}}{(\xi+\frac{y_2}{h})^2}\left[\frac{1}{\xi-\frac{\sqrt{y_2}}{h}}+\frac{1}{\xi+\frac{\sqrt{y_2}}{h}}-\frac{2}{\xi+\frac{y_2}{h}}\right]\,d\xi,\end{aligned} \tag{3.28}$$

and

$$\begin{aligned}cov(X_{f_i},X_{f_j}) = -\lim_{r\downarrow 1}\frac{\kappa}{4\pi^2}\oint_{|\xi_1|=1}\oint_{|\xi_2|=1}\frac{f_i\left(\frac{1+h^2+2h\Re(\xi_1)}{(1-y_2)^2}\right)f_j\left(\frac{1+h^2+2h\Re(\xi_2)}{(1-y_2)^2}\right)}{(\xi_1-r\xi_2)^2}\,d\xi_1 d\xi_2\\ -\frac{(\beta_x y_1+\beta_y y_2)(1-y_2)^2}{4\pi^2h^2}\oint_{|\xi_1|=1}\frac{f_i\left(\frac{1+h^2+2h\Re(\xi_1)}{(1-y_2)^2}\right)}{(\xi_1+\frac{y_2}{h})^2}d\xi_1\oint_{|\xi_2|=1}\frac{f_j\left(\frac{1+h^2+2h\Re(\xi_2)}{(1-y_2)^2}\right)}{(\xi_2+\frac{y_2}{h})^2}d\xi_2.\end{aligned} \tag{3.29}$$

The examples below describe applications of Theorem 3.10 to some important linear spectral statistics of the Fisher matrix $\mathbf{F}_n$. The results are derived using Proposition 3.11 and contour integral along the same lines as in the calculations given in the proof of Proposition 3.8. The details are left to the reader.

Example 3.12 For $f_1 = \log(a+bx)$, $f_2 = \log(a'+b'x)$, $a,\ a' \geq 0$, $b,\ b' > 0$, we have for the real case ($\kappa = 2$),

$$\mathbb{E}X_{f_1} = \frac{1}{2}\log\left(\frac{(c^2-d^2)h^2}{(ch-y_2d)^2}\right)-\frac{\beta_x y_1(1-y_2)^2d^2}{2\,(ch-dy_2)^2}+\frac{\beta_y(1-y_2)}{2}\left[\frac{2dy_2}{ch-dy_2}+\frac{d^2\left(y_2^2-y_2\right)}{(ch-dy_2)^2}\right]$$

and

$$\mathrm{cov}(X_{f_1},X_{f_2}) = 2\log\left(\frac{cc'}{cc'-dd'}\right)+\frac{(\beta_x y_1+\beta_y y_2)(1-y_2)^2dd'}{(ch-dy_2)(c'h-d'y_2)}$$

where $c > d > 0$, $c' > d' > 0$ satisfying $c^2+d^2 = \frac{a(1-y_2)^2+b(1+h^2)}{(1-y_2)^2}$, $(c')^2+(d')^2 = \frac{a'(1-y_2)^2+b'(1+h^2)}{(1-y_2)^2}$, $cd = \frac{bh}{(1-y_2)^2}$ and $c'd' = \frac{b'h}{(1-y_2)^2}$.

Example 3.13 For $g_k(x) = x^k$ and $g_l(x) = x^l$ with positive integers $k \geq l \geq 1$, we have in the real case ($\kappa = 2$),

$$\mathbb{E}X_{g_k} = \frac{1}{2(1-y_2)^{2k}}\left[(1-h)^{2k}+(1+h)^{2k}-2(1-y_2)^k\left(1-\frac{h^2}{y_2}\right)^k\right.$$

$$+\sum_{i_1+i_2+i_3=k-1}\frac{k\cdot k!i_3!\left((-1)^{i_3}+(-1)^{2i_3+1}\right)}{(k-i_1)!(k-i_2)!}h^{k+i_1-i_2}+\sum_{i_1+i_2+i_3=k-1}\frac{2k\cdot k!i_3!\cdot h^{k+i_1-i_2}}{(k-i_1)!(k-i_2)!}\left(-\frac{h}{y_2}\right)^{i_3+1}\Bigg]$$

$$+\frac{\beta_x\cdot y_1}{h^2(1-y_2)^{2(k-1)}}\Bigg[\sum_{i_1+i_2+i_3=2}\frac{k\cdot k!(k+i_3-1)!(-1)^{i_3}}{2(k-i_1)!(k-i_2)!}h^{i_1}(1-y_2)^{k-i_1}\left(\frac{h^2-y_2}{h}\right)^{k-i_2}\left(-\frac{h}{y_2}\right)^{k+i_3}$$

$$+\sum_{i_1+i_2+i_3=k-1}\frac{k\cdot k!(2+i_3)!(-1)^{i_3}}{(k-i_1)!(k-i_2)!2!}\cdot h^{k+i_1-i_2}\left(\frac{h}{y_2}\right)^{3+i_3}\Bigg]$$

$$+\frac{\beta_y}{2(1-y_2)^{2k-1}}\Bigg[\sum_{\substack{i_1+i_2\\+i_3+i_4=1}}\frac{k\cdot k!(k+i_4-1)!(-1)^{i_4}}{(k-i_1)!(k-i_2)!}h^{i_1}(1-y_2)^{k-i_1}\left(\frac{h^2-y_2}{h}\right)^{k-i_2}\left(\frac{\sqrt{y_2}-y_2}{h}\right)^{1-i_3}\left(-\frac{h}{y_2}\right)^{k+i_4}$$

$$+\sum_{i_1+i_2+i_3+i_4=1}\frac{k\cdot k!(k+i_4-1)!(-1)^{i_4}}{(k-i_1)!(k-i_2)!}h^{i_1}(1-y_2)^{k-i_1}\left(\frac{h^2-y_2}{h}\right)^{k-i_2}\left(\frac{-\sqrt{y_2}-y_2}{h}\right)^{1-i_3}\left(-\frac{h}{y_2}\right)^{k+i_4}$$

$$+\sum_{i_1+i_2+i_3+i_4=k-1}\frac{k\cdot k!(i_4+1)!(-1)^{i_4}}{(k-i_1)!(k-i_2)!}\cdot h^{k+i_1-i_2}\left(\left(\frac{\sqrt{y_2}}{h}\right)^{1-i_3}+\left(-\frac{\sqrt{y_2}}{h}\right)^{1-i_3}\right)\left(\frac{h}{y_2}\right)^{2+i_4}$$

$$-\sum_{\substack{i_1+i_2+i_3\\+i_4+i_5=2}}\frac{k\cdot k!(k+i_5-1)!h^{i_1}(1-y_2)^{k-i_1}}{(k-i_1)!(k-i_2)!(-1)^{i_5}}\left(\frac{h^2-y_2}{h}\right)^{k-i_2}\left(\frac{\sqrt{y_2}-y_2}{h}\right)^{1-i_3}\left(\frac{-\sqrt{y_2}-y_2}{h}\right)^{1-i_4}\left(-\frac{h}{y_2}\right)^{k+i_5}$$

$$-\sum_{i_1+i_2+i_3+i_4+i_5=k-1}\frac{k\cdot k!(i_5+2)!(-1)^{i_5}}{(k-i_1)!(k-i_2)!}\cdot h^{k+i_1-i_2}\left(\frac{\sqrt{y_2}}{h}\right)^{1-i_3}\left(-\frac{\sqrt{y_2}}{h}\right)^{1-i_4}\left(\frac{h}{y_2}\right)^{3+i_5}\Bigg]$$

and

$$\begin{aligned}
&\operatorname{cov}(X_{g_k},X_{g_l})\\
&=\frac{2}{(1-y_2)^{2l+2k}}\sum_{\substack{i_1+i_2+i_3\\=l-1}}\sum_{\substack{j_1+j_2\\=k+i_3+1}}\frac{l\cdot l!(i_3+1)!k!k!}{(l-i_1)!(l-i_2)!(k+i_3+1)!(k-j_1)!(k-j_2)!}h^{l+k+i_1-i_2+j_1-j_2}\\
&+\frac{(\beta_xy_1+\beta_yy_2)(1-y_2)^2}{h^2}\Bigg\{\Bigg[\sum_{i_1+i_2+i_3=1}\frac{l\cdot l!(l+i_3-1)!(-1)^{i_3}}{(l-i_1)!(l-i_2)!}h^{i_1}(1-y_2)^{l-i_1}\left(\frac{h^2-y_2}{h}\right)^{l-i_2}\left(-\frac{h}{y_2}\right)^{l+i_3}\\
&\qquad+\sum_{i_1+i_2+i_3=l-1}\frac{l\cdot l!(1+i_3)!(-1)^{i_3}}{(l-i_1)!(l-i_2)!}h^{l+i_1-i_2}\left(\frac{h}{y_2}\right)^{2+i_3}\Bigg]\\
&\times\Bigg[\sum_{i_1+i_2+i_3=1}\frac{k\cdot k!(k+i_3-1)!(-1)^{i_3}}{(k-i_1)!(k-i_2)!}h^{i_1}(1-y_2)^{k-i_1}\left(\frac{h^2-y_2}{h}\right)^{k-i_2}\left(-\frac{h}{y_2}\right)^{k+i_3}\\
&\qquad+\sum_{i_1+i_2+i_3=k-1}\frac{k\cdot k!(1+i_3)!(-1)^{i_3}}{(k-i_1)!(k-i_2)!}h^{k+i_1-i_2}\left(\frac{h}{y_2}\right)^{2+i_3}\Bigg]\Bigg\}
\end{aligned}$$

Example 3.14 If $g=e^x$, then by Taylor expansion, we have

$$\mathbb{E}X_g=\sum_{l=0}^{+\infty}\frac{1}{l!}\mathbb{E}X_{g_l}\quad\text{and}\quad\operatorname{cov}(X_f,X_f)=\sum_{k,l=0}^{+\infty}\operatorname{cov}(X_{g_k},X_{g_l})$$

where $g_l(x)=x^l$, $\mathbb{E}X_{g_l}$ and $\operatorname{cov}(X_{g_l},X_{g_k})$are given in Example 3.13.

3.5 The substitution principle

So far we have studied the *non-centred* sample covariance (2.7)

$$\mathbf{S}_{n0} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{i0}\mathbf{x}_{i0}{}^{*} , \tag{3.30}$$

from a sample $\mathbf{x}_{10}, \ldots, \mathbf{x}_{n0}$ of a p-dimensional population. The 0 in the subscript here is used to remind the fact that so far it has been assumed that the population is centred, i.e. $\mathbb{E}\,\mathbf{x}_{i0} = \mathbf{0}$. However, in real-life statistical applications, populations have in general a non null mean $\boldsymbol{\mu}$. If the sample is denoted as $\mathbf{x}_1, \ldots, \mathbf{x}_n$, the *centred* sample covariance matrix in (2.1)

$$\mathbf{S}_n = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^* , \tag{3.31}$$

is preferred ($\overline{\mathbf{x}} = \frac{1}{n}\sum_j \mathbf{x}_j$ is the sample mean). Recall that the population covariance matrix is $\boldsymbol{\Sigma} = \mathbf{I}_p$ in both situations.

Is there then any difference between these centred and non-centred sample covariance matrices regarding their eigenvalue statistics? Consider first the limiting spectral distributions. Let $\lambda_{10} \geq \cdots \geq \lambda_{p0}$ and $\lambda_1 \geq \cdots \geq \lambda_p$ be the ordered eigenvalues of $\mathbf{S}_n^0$ and $\mathbf{S}_n$, respectively. Write

$$\mathbf{S}_n = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^* - \frac{n}{n-1}(\boldsymbol{\mu} - \overline{\mathbf{x}})(\boldsymbol{\mu} - \overline{\mathbf{x}})^* .$$

The first sum is distributed as $\frac{n}{n-1}\mathbf{S}_{n0}$ while the second term is a random matrix of rank one (notice however the division by $n-1$ in the first term). By Cauchy interlacing theorem, we have

$$\frac{n}{n-1}\lambda_{10} \geq \lambda_1 \geq \frac{n}{n-1}\lambda_{20} \geq \lambda_2 \geq \cdots \geq \frac{n}{n-1}\lambda_{p0} \geq \lambda_p .$$

It follows that the ESD's $F^{\mathbf{S}_{n0}}$ and $F^{\mathbf{S}_n}$ of the two matrices converge to a same LSD, namely the Marčenko-Pastur distribution F_y with index $y = \lim p/n$.

Next consider the fluctuations of a linear spectral statistics from the two matrices. Let g be a smooth function. By Theorem 3.4

$$p[F^{\mathbf{S}_{n0}}(g) - F_{y_n}(g)] = g(\lambda_{10}) + \cdots + g(\lambda_{p0}) - pF_{y_n}(g) \xrightarrow{\mathscr{D}} \mathcal{N}(m(g), v(g)) , \tag{3.32}$$

a Gaussian distribution whose parameters $m(g)$ and $v(g)$ depend only on the M-P law F_y and g. Is this also true for the sample covariance matrix $\mathbf{S}_n$, namely

$$p[F^{\mathbf{S}_n}(g) - F_{y_n}(g)] = g(\lambda_1) + \cdots + g(\lambda_p) - pF_{y_n}(g) \xrightarrow{\mathscr{D}} \mathcal{N}(m(g), v(g)) , \tag{3.33}$$

with the same limiting parameters $(m(g), v(g))$?

The crucial issue here is that the centring term $pF_{y_n}(g)$ uses a finite-horizon proxy of the LSD F_y obtained by substituting the current dimension-to-sample ratio $y_n = p/n$ for its limit y. Since p is of the order of n, any mis-estimation of order n^{-1} in $F_{y_n}(g)$ will affect the asymptotic mean $m(g)$.

It turns out that linear spectral statistics of $\mathbf{S}_n$ and $\mathbf{S}_n^0$ do not share a same CLT, that is the convergence in (3.33) is not true as such. This can be best explained by observing the Gaussian case. Define $N = n-1$ to be the *adjusted sample size*. For a Gaussian population,

$N\mathbf{S}_n := \sum_{i=1}^n (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^*$ has a Wishart distribution $\mathbf{W}_N$ with N degrees of freedom. Since from a centred Gaussian population, the matrix $N\mathbf{S}_N^0 = \sum_{i=1}^N \mathbf{x}_{i0}\mathbf{x}_{i0}{}^*$ has the same Wishart distribution, we conclude that the fluctuations of the eigenvalues (λ_j) of $\mathbf{S}_n$ are the same as the matrix $\mathbf{S}_N^0$ so that by (3.32), it holds

$$p\left\{F^{\mathbf{S}_n}(g) - pF_{y_N}(g)\right\} \xrightarrow{\mathscr{D}} \mathcal{N}(m(g), v(g))\ . \tag{3.34}$$

In words, in the Gaussian case, the CLT for the centred sample covariance matrix is the same as the CLT for the non-centred sample covariance matrix provided that in the centring parameter one substitutes the adjusted sample size $N = n - 1$ for the sample size n. This result will be referred as the *substitution principle*. Notice that typically the difference between $F_{y_N}(g)$ and $F_{y_n}(g)$ is of order n^{-1} and as explained above, such a difference is non negligible because of the multiplication by p in the CLT.

Example 3.15 For $y < 1$ and $g(x) = \log x$, Example 2.11 shows that $F^y(g) = -1 + (y - 1)\log(1 - y)/y$. Therefore

$$F_{y_n}(g) - F_{y_N}(g) = -\frac{1}{n}\left\{1 + \frac{1}{y_n}\log(1 - y_n)\right\} + o(n^{-1})\ ,$$

so that

$$p\left\{F_{y_n}(g) - F_{y_N}(g)\right\} \to -\{y + \log(1 - y)\} > 0, \quad \text{as } n \to \infty.$$

So using N or n in the centring parameter of the CLT leads to a different asymptotic mean $m(g)$.

This substitution principle is indeed a remarkable result and provides an elegant solution to the question raised in (3.33). It then raises the question whether the principle is universal, i.e. valid for general populations other than Gaussian. The following theorem establishes this universality.

Theorem 3.16 *(One sample substitution principle) Assume the same conditions as in Theorem 3.4 except that the zero mean condition $\mathbb{E}\, x_{ij} = 0$ is dropped and the sample covariance matrix $\mathbf{S}_n$ is defined as in 3.31. Then, with $N = n - 1$ and $y_N = p/N$, the random vector $\{X_n(f_1), \cdots X_n(f_k)\}$ where*

$$X_n(f) = p\left\{F^{\mathbf{S}_n}(f) - F_{y_N}(f)\right\}$$

converges weakly to the same Gaussian vector $(X_{f_1}, \cdots X_{f_k})$ given in Theorem 3.4.

Next consider the two-sample Fisher matrix $\mathbf{F}_{n0} = \mathbf{S}_{10}\mathbf{S}_{20}^{-1}$ in (2.24). Again the subscript 0 is added to remind the fact that both populations have zero mean. When these means are non null, it is more natural to consider centred sample covariance matrices and the associated Fisher matrix $\mathbf{F}_n$ defined by

$$\mathbf{S}_1 = \frac{1}{N_1}\sum_{k=1}^{n_1}(\mathbf{x}_k - \overline{\mathbf{x}})(\mathbf{x}_k - \overline{\mathbf{x}})^*, \quad \mathbf{S}_2 = \frac{1}{N_2}\sum_{k=1}^{n_2}(\mathbf{y}_k - \overline{\mathbf{y}})(\mathbf{y}_k - \overline{\mathbf{y}})^*, \quad \mathbf{F}_n = \mathbf{S}_1\mathbf{S}_2{}^{-1}, \tag{3.35}$$

where $N_i = n_i - 1$, $i = 1, 2$ are the adjusted sample sizes, and $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$ the sample means from the two samples. Above discussions on the sample covariance matrix reveal that in this case, the CLT for linear spectral statistics of a Fisher random matrix could be different of the CLT given in Theorem 3.10 for zero-mean populations. Again, considering

Gaussian populations indicates that there might be a substitution principle. Such universal principle indeed exists.

Theorem 3.17 *(Two-sample substitution principle) Assume the same conditions as in Theorem 3.10 except that the zero mean conditions* $\mathbb{E}\,x_{ij} = 0$ *and* $\mathbb{E}\,y_{ij} = 0$ *are dropped, and the sample covariance matrix and the associated Fisher matrix* $\mathbf{F}_n$ *are defined as in* (3.35)*. Then, with* $N_i = n_i - 1$ *and* $y_{N_i} = p/N_i$*,* $i = 1, 2$*, the random vector* $\{X_n(f_1), \cdots X_n(f_k)\}$ *where*

$$X_n(f) = p\left\{F^{\mathbf{F}_n}(f) - F_{y_{N_1}, y_{N_2}}(f)\right\}$$

converges weakly to the same Gaussian vector $(X_{f_1}, \cdots X_{f_k})$ *given in Theorem 3.10.*

To summarise, for any statistical application developed later in this book that uses a CLT for some linear spectral statistic of a sample covariance matrix or of a Fisher matrix, it is sufficient to indicate the results for zero-mean populations with the covariance matrix $\mathbf{s}_{n0}$ or $\mathbf{F}_{n0}$. The corresponding results for populations with unknown means are easily derived using the substitution principles of this section.

Notes

Central limit theorems for eigenvalues of sample covariance matrices have a long history. The earliest work dates back to Jonsson (1982) for Gaussian samples. The breakthrough work on the topic is due to Bai and Silverstein (2004) for general samples by providing explicit expressions for the limiting mean and variance functions. This is the CLT presented in §3.3 However, this CLT requires that the first four moments of the sample variables match those of the Gaussian case.

Recent efforts have been made in Pan and Zhou (2008) and Lytova and Pastur (2009) to overcome these moment restrictions. Theorem 3.4 is an adaptation of the CLT in Pan and Zhou (2008) to the present case. The representation of the limiting parameters using contour integrals on the unit circle is due to Wang and Yao (2013).

The CLT for linear spectral statistics of random Fisher matrix is due to Zheng (2012).

In both Theorems 3.4 and 3.10, the random variables are assumed to be independent and identically distributed. The assumption of identical distribution can be removed by imposing a Lindeberg condition on the moments of the independent variables, see e.g. Bai and Silverstein (2010) for an approach along this line.

The substitution principles in Theorems 3.16 and 3.17 are due to Zheng et al. (2015). An earlier and closely related result for the sample covariance appeared in Pan (2014). In this reference, the sample covariance matrix is normalised by $1/n$ and the proposed solution is to find a direct correction to the asymptotic mean. Such correction is unnecessary in the substitution principle with the normalisation $1/N$.

4

The generalised variance and multiple correlation coefficient

4.1 Introduction

For linear spectral statistics of a large sample covariance matrix or of a Fisher matrix, their limiting values and limiting distributions are derived and discussed in Chapter 2 and Chapter 3, respectively. This chapter is devoted to applications of this general theory to two traditional multivariate statistics, namely the *generalised variance* and the *multiple correlation coefficient*. Despite their relative simplicity, the two applications nicely illustrate the whole methodology developed in the book. The main message is that with the help of the new theory, it is possible to find an asymptotic framework capable to deal with large-dimensional data. In particular, new limiting distributions derived within this framework for traditional multivariate statistics provide accurate finite-sample approximations in case of large-dimensional data. More sophisticated applications and examples are developed in later chapters of the book.

4.2 The generalised variance

The variance σ^2 of a univariate distribution has two multivariate analogues, namely the covariance matrix $\mathbf{\Sigma}$ and the scalar $|\mathbf{\Sigma}|$. The later is called *the generalised variance* of the multivariate distribution. Similarly, the generalised variance of the sample of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is

$$|\mathbf{S}| = \left| \frac{1}{N} \sum_{k=1}^{n} (\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})' \right|, \tag{4.1}$$

where $N := n - 1$ is the degree of freedom. In some sense each of these is a measure of spread. The generalised variance is important for multivariate analysis since it occurs in many likelihood ratio criteria for testing hypotheses.

Before going further, we introduce a very important class of distributions, namely the Wishart distributions.

Definition 4.1 Let $\mathbf{x}_k$, $k = 1, \ldots, n$ be i.i.d. from a p-th dimensional normal distribution $\mathcal{N}_p(\boldsymbol{\mu}_k, \mathbf{\Sigma})$. Then the distribution of the $p \times p$ random matrix

$$\mathbf{W} = \sum_{k=1}^{n} \mathbf{x}_k \mathbf{x}_k' \tag{4.2}$$

is called a Wishart distribution with degrees of freedom n and non-central matric parameter $\mathbf{\Psi} = \sum_{k=1}^{n} \mathbf{\Sigma}^{-1/2}\boldsymbol{\mu}_k\boldsymbol{\mu}_k'\mathbf{\Sigma}^{-1/2}$. The distribution is denoted by $W(\mathbf{\Sigma}, n, \mathbf{\Psi})$

Moreover, when $\mathbf{\Psi} = 0$, the distribution is a centred Wishart distribution and denoted by $W(\mathbf{\Sigma}, n)$. In particular, $W(\mathbf{I}_p, n)$ denotes the *standard* Wishart distribution with n degrees of freedom.

Fundamental properties such as the density function, moments and the characteristic function are well known; we refer the reader to Anderson (2003, Chapter 7) for a complete account.

4.2.1 Exact distribution with a normal sample

Consider a sample $\mathbf{x}_k$, $k = 1, \ldots, n$ from a pth dimensional normal distribution $\mathcal{N}(\boldsymbol{\mu}_k, \mathbf{\Sigma})$. The generalised variance $|\mathbf{S}|$ in (4.1) has the same distribution as $|\mathbf{A}/N|$, where $\mathbf{A} = \sum_{k=1}^{N} \boldsymbol{\xi}_k\boldsymbol{\xi}_k'$, and the $\boldsymbol{\xi}_k$s are i.i.d. random vectors distributed as $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. (Recall $N = n - 1$). Write $\boldsymbol{\xi}_k = \mathbf{B}\mathbf{u}_k, k = 1, \ldots, N$, where $\mathbf{B}$ is invertible and $\mathbf{B}\mathbf{B}' = \mathbf{\Sigma}$. It follows that $\mathbf{u}_1, \ldots, \mathbf{u}_N$ are i.i.d. distributed as $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

Let

$$\mathbf{M} = \sum_{k=1}^{N} \mathbf{u}_k\mathbf{u}_k' = \sum_{k=1}^{N} \mathbf{B}^{-1}\boldsymbol{\xi}_k\boldsymbol{\xi}_k'(\mathbf{B}^{-1})' = \mathbf{B}^{-1}\mathbf{A}\mathbf{B}^{-1} \ ;$$

then $|\mathbf{A}| = |\mathbf{B}| \cdot |\mathbf{M}| \cdot |\mathbf{B}'| = |\mathbf{M}| \cdot |\mathbf{\Sigma}|$. Note that $\mathbf{M}$ is distributed as the p-dimensional standard Wishart distribution $W(\mathbf{I}_p, N)$ with N degrees of freedom. It is well-known that its determinant $|\mathbf{M}|$ is distributed as a product of independent chi-squared random variables.

Theorem 4.2 *The generalised variance $|\mathbf{S}|$ of a sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$, where the $\{\mathbf{x}_i\}$ are normally distributed with mean $\boldsymbol{\mu}$ and variance $\mathbf{\Sigma}$, has the same distribution as $N^{-p}|\mathbf{\Sigma}|$ multiplied by the product of p independent factors, say u_k, $i = k, \ldots, p$, the factor u_k having the chi-squared distribution with $N - k + 1$ degrees of freedom.*

For a proof of this classical result, we refer the reader to Anderson (2003, §7.2). For example when $p = 1$, $|\mathbf{S}|$ has the distribution of $|\mathbf{\Sigma}| \cdot \chi_N^2/N$. If $p = 2$, $|\mathbf{S}|$ has the distribution of $|\mathbf{\Sigma}|\chi_N^2 \cdot \chi_{N-1}^2/N^2$. The general result from the theorem can be expressed by the distributional identity

$$|\mathbf{S}| \overset{\mathscr{D}}{=} N^{-p}|\mathbf{\Sigma}| \cdot \chi_N^2 \cdot \chi_{N-1}^2 \cdots \chi_{N-p+1}^2,$$

where the χ^2s are independent.

4.2.2 Large sample distribution from a normal population

Rewrite the distributional identity in Theorem 4.2 as

$$\frac{|\mathbf{S}|}{|\mathbf{\Sigma}|} \overset{\mathscr{D}}{=} \prod_{k=1}^{p} \frac{\chi_{N-k+1}^2}{N}.$$

To find large sample limiting distribution, let p be fixed and $N \to \infty$. For each $1 \leq k \leq p$, $[\chi_{N-k+1}^2 - (N - k + 1)]/\sqrt{N - k + 1} \xrightarrow{\mathscr{D}} \mathcal{N}(0, 2)$; or equivalently $\sqrt{N}(\chi_{N-k+1}^2/N - 1) \xrightarrow{\mathscr{D}}$

$\mathcal{N}(0,2)$. Since these chi-squares are independent, the vector

$$\mathbf{y} = \frac{1}{N}\begin{pmatrix} \chi^2_N \\ \vdots \\ \chi^2_{N-p+1} \end{pmatrix},$$

is asymptotic normal: $\sqrt{N}(\mathbf{y} - \mathbf{1}_p) \xrightarrow{\mathscr{D}} \mathcal{N}(\mathbf{0}, 2\mathbf{I}_p)$ Consider the function $f : \mathbb{R}^p \to \mathbb{R}$; $\mathbf{u} = (u_1, \ldots, u_p) \mapsto u_1 u_2 \cdots u_p$. Then $|\mathbf{S}|/|\mathbf{\Sigma}| = f(\mathbf{y})$, $f(\mathbf{1}_p) = 1$, and $\nabla f(\mathbf{1}_p) = \mathbf{1}'_p$. Applying the delta method,

$$\sqrt{N}\left(\frac{|\mathbf{S}|}{|\mathbf{\Sigma}|} - 1\right)$$

converges in distribution to $\mathcal{N}(0, 2p)$ when $N \to \infty$.

Theorem 4.3 *For the sample covariance matrix* $\mathbf{S}$ *from a p-dimensional normal population* $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$*,* $\sqrt{n}(|\mathbf{S}|/|\mathbf{\Sigma}| - 1)$ *converges in distribution to* $\mathcal{N}(0, 2p)$ *when* $N \to \infty$ *(while the dimension p is hold fixed).*

4.2.3 The generalised variance from a large-dimensional sample

Clearly, when the dimension p is larger than the sample size n, the sample covariance matrix $\mathbf{S}$ is singular. The sample generalised variance is null and cannot be a reliable estimate of the population generalised variance. What happens for dimension p smaller than n? And does the sample generalised variance approach its population counterpart for large sample sizes?

We start with a normal population $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_p)$ and assume that $p/n \to y \in (0,1)$. Define for $u \in (0,1)$,

$$d(u) = 1 + \frac{1-u}{u}\log(1-u) = \sum_{k=1}^{\infty} \frac{1}{k(k+1)} u^k, \tag{4.3}$$

which is a positive function. For a standard normal population, the generalised variance is unit. For the sample generalised variance, consider

$$\frac{1}{p}\log|\mathbf{S}| = \int_0^{\infty} \log x dF^{\mathbf{S}}(x).$$

By the Marčenko-Pastur law, almost surely, $F^{\mathbf{S}} \xrightarrow{\mathscr{D}} F_y(x)$, where F_y is the Marčenko-Pastur distribution with index y and scale parameter $\sigma^2 = 1$, see §2.3, Eq.(2.5). Furthermore, by Theorem 6.1, almost surely, $\lambda_1 \to b = (1+\sqrt{y})^2$ and $\lambda_p \to a = (1-\sqrt{y})^2$. By Helly-Bray's theorem,

$$\frac{1}{p}\log|\mathbf{S}| \xrightarrow{\text{a.s.}} \int_a^b \log x dF_y(x) = \int_a^b \frac{\log x}{2\pi x y}\sqrt{(x-a)(b-x)}dx. \tag{4.4}$$

As proved in Example 2.11, the last integral equals $-d(y)$.

Theorem 4.4 *Under the large-dimensional scheme* $p \sim n$ *with* $p/n \to y \in (0,1)$ *and for a normal population* $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_p)$*,* $p^{-1}\log|\mathbf{S}|$ *converges almost surely to* $-d(y)$ *(the function d is defined in* (4.3)*).*

When the population covariance matrix changes from $\mathbf{I}_p$ to $\boldsymbol{\Sigma}$, the sample generalised variance is multiplied by $|\boldsymbol{\Sigma}|$. We get the following theorem.

Theorem 4.5 *Under the large-dimensional scheme $p \sim n$ with $p/n \to y \in (0, 1)$ and for a normal population $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we have*

$$\frac{1}{p} \log(|\mathbf{S}|/|\boldsymbol{\Sigma}|) \xrightarrow{a.s.} -d(y). \tag{4.5}$$

Therefore, for large-dimensional data, the sample generalised variance *is not* a consistent estimator of the population generalised variance and in general it has a negative bias.

It is also worth mentioning that, by Theorems 2.14 and 6.1, Theorems 4.5 and 4.5 are still valid for non normal samples provided that the population distribution has a finite fourth moment.

We now give a central limit theorem for the sample generalised variance. In Theorem 4.3, we have proved that for fixed p and $n \to \infty$,

$$\sqrt{n/2p}(|\mathbf{S}|/|\boldsymbol{\Sigma}| - 1) \xrightarrow{\mathscr{D}} \mathcal{N}(0, 1).$$

This result can be used to test hypotheses on the population generalised variance $|\boldsymbol{\Sigma}|$. However, Theorem4.5 indicates that under large-dimensional scheme, such procedures may suffer from severe inconsistency. The following theorem provides a new central limit theorem.

Theorem 4.6 *Under the conditions of Theorem 4.5, we have*

$$\log(|\mathbf{S}|/|\boldsymbol{\Sigma}|) + pd(y_n) \xrightarrow{\mathscr{D}} \mathcal{N}(\mu, \sigma^2), \tag{4.6}$$

where $y_n = p/N$ and

$$\mu = \frac{1}{y} \log(1 - y), \quad \sigma^2 = -2 \log(1 - y).$$

One should notice that the centring term in the above theorem depends on the sample size n (through y_n). This is a common feature for large sample covariance matrices since the convergence of $y_n \to y$ can be arbitrarily slow. There is then no way to use a centring term independent from n. Moreover, from the point of view of application, we know y_n only and the limit y exists only virtually. This means that for the calculation of any parameter involving in an asymptotic limit or distribution, we need always to substitute y_n for the theoretical y.

The proof of Theorem 4.6 is a simple application of the general central limit theorem 3.4 and is left to the reader.

4.2.4 Hypothesis testing and confidence intervals for the generalised variance

In a straightforward manner, the above central limit theorem can be used for testing hypotheses about the generalised variance. To test

$$H_0 : |\boldsymbol{\Sigma}| = a_0, \ \text{ v.s. } \ H_1 : |\boldsymbol{\Sigma}| \neq a_0,$$

by using Theorem 4.6, we reject the null hypothesis when

$$\left|\log(|\mathbf{S}|/a_0) + pd(y_n) - \frac{1}{y_n}\log(1-y_n)\right| > z_{\alpha/2}\sqrt{-2\log(1-y_n)}.$$

We denote this large-dimensional procedure by [L].

If we use the traditional central limit theorem 4.3, we will reject the null hypothesis when

$$\left|\frac{|\mathbf{S}|}{a_0} - 1\right| > z_{\alpha/2}\sqrt{2p/n}.$$

This traditional procedure is referred as [C].

Similarly, using these two central limit theorems, we can design one-sided tests for alternative hypotheses $H_1 : |\mathbf{\Sigma}| < a_0$ and $H_1 : |\mathbf{\Sigma}| > a_0$, respectively.

So, which one of the two procedures is better? To answer the question, we conduct a Monte-Carlo experiment to compare the size and the power of these two procedures. Data are sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ under the null hypothesis and from $\mathcal{N}(\mathbf{0}, 0.95\mathbf{I}_p + 0.05\mathbf{1}_p\mathbf{1}_p')$ under the alternative hypothesis. The number of independent replications is 10000. Two sample sizes $n = 500$ and $n = 1000$ are combined with 5 dimensions $p \in \{5, 10, 50, 100, 300\}$. All the three alternative hypotheses are examined (i.e. two-sided, one-sided lower and one-sided upper). Empirical results for procedure [L] and [C] are reported in Tables 4.1 and 4.2, respectively.

Table 4.1 *Empirical size and power of tests derived from Theorem 4.6.*

	Size			Power		
	2-sided	1-sided L	1-sided U	2-sided	1-sided L	1-sided U
(p = 300)	0.0513	0.0508	0.0528	1.0	1.0	0.0
(p = 100)	0.0516	0.0514	0.0499	0.997	1.0	0.0
(p = 50)	0.0488	0.0471	0.0504	0.785	0.866	0.0
(p = 10)	0.0507	0.0524	0.0489	0.0732	0.117	0.0168
(p = 5)	0.0507	0.0517	0.0497	0.050	0.0695	0.0331
(n = 500)						
(p = 300)	0.0496	0.0496	0.0493	1.0	1.0	0.0
(p = 100)	0.0508	0.0509	0.0515	1.0	1.0	0.0
(p = 50)	0.0523	0.0501	0.0517	0.979	0.990	0.0
(p = 10)	0.0506	0.0498	0.0504	0.0969	0.1591	0.0116
(p = 5)	0.0508	0.0530	0.0494	0.0542	0.0784	0.0288
(n = 1000)						

These results can be summarised as follows. The traditional procedure [C] becomes quickly inconsistent when the dimension p increases: for dimensions exceeding 50, its size is almost 1 and even for low dimensions such as 5 or 10, the size (two-sided test and one-sided lower test) is higher than the nominal one (indeed the test statistic has a positive and diverging drift). By contrast, the large-dimension procedure [L] is consistent as expected for large dimensions (e.g. 100 and 300). Moreover and what is really surprising, even for moderate or low dimensions such as 5 or 10, the empirical sizes of [L] remain almost always better than the traditional procedure [C]. Therefore, one should use the large-dimensional corrected procedure [L] even for low-dimensional data.

Table 4.2 *Empirical size and power of tests derived from Theorem 4.3.*

	Size			Power		
	2-sided	1-sided L	1-sided U	2-sided	1-sided L	1-sided U
$(p = 300)$	1.0	1.0	0.0	1.0	1.0	0.0
$(p = 100)$	1.0	1.0	0.0	1.0	1.0	0.0
$(p = 50)$	1.0	1.0	0.0	1.0	1.0	0.0
$(p = 10)$	0.09	0.14	0.014	0.17	0.26	0.0023
$(p = 5)$	0.057	0.078	0.031	0.065	0.10	0.019
			$(n = 500)$			
$(p = 300)$	1.0	1.0	0.0	1.0	1.0	0.0
$(p = 100)$	1.0	1.0	0.0	1.0	1.0	0.0
$(p = 50)$	0.9817	0.9918	0.0	1.0	1.0	0.0
$(p = 10)$	0.0666	0.1067	0.0209	0.1801	0.2623	0.0037
$(p = 5)$	0.0530	0.0696	0.0360	0.0664	0.1040	0.0203
			$(n = 1000)$			

Lastly, using the well-known relationship between critical regions of test and confidence intervals, we find a two-sided confidence interval with (asymptotic) level $(1 - \alpha)$ for the generalised variance:

$$|\mathbf{\Sigma}| \in |\mathbf{S}| \exp\left\{ pd(y_n) - \frac{1}{y_n}\log(1 - y_n) \pm z_{\alpha/2}\sqrt{-2\log(1 - y_n)} \right\}.$$

4.3 The multiple correlation coefficient

Consider a p-dimensional population $\mathbf{x} = (X_1, X_2, \ldots, X_p)$ with covariance matrix $\text{cov}(\mathbf{x}) = \mathbf{\Sigma}$. The *multiple correlation coefficient* between one variable X_1 and the vector $\mathbf{X}_2 = (X_2, \ldots, X_p)'$ in the population is

$$\bar{R} = \frac{\boldsymbol{\beta}'\mathbf{\Sigma}_{22}\boldsymbol{\beta}}{\sqrt{\sigma_{11}\boldsymbol{\beta}'\mathbf{\Sigma}_{22}\boldsymbol{\beta}}} = \sqrt{\frac{\boldsymbol{\beta}'\mathbf{\Sigma}_{22}\boldsymbol{\beta}}{\sigma_{11}}} = \sqrt{\frac{\boldsymbol{\sigma}_1'\mathbf{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_1}{\sigma_{11}}},$$

where $\boldsymbol{\beta}$, $\boldsymbol{\sigma}_1$ $\mathbf{\Sigma}_{22}$ are defined by

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & \boldsymbol{\sigma}_1' \\ \boldsymbol{\sigma}_1 & \mathbf{\Sigma}_{22} \end{pmatrix}, \quad \boldsymbol{\beta} = \mathbf{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_1.$$

Given a sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ $(n > p)$, we estimate $\mathbf{\Sigma}$ by $\mathbf{S} = [n/N]\hat{\mathbf{\Sigma}}$ or

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{A} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\mathbf{x}})(\mathbf{x}_k - \hat{\mathbf{x}})' = \begin{pmatrix} \hat{\sigma}_{11} & \hat{\boldsymbol{\sigma}}_1' \\ \hat{\boldsymbol{\sigma}}_1 & \hat{\mathbf{\Sigma}}_{22} \end{pmatrix} = \frac{1}{n}\begin{pmatrix} a_{11} & \mathbf{a}_1^* \\ \mathbf{a}_1 & \mathbf{A}_{22} \end{pmatrix}.$$

and we estimate $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}} = \hat{\mathbf{\Sigma}}_{22}^{-1}\hat{\boldsymbol{\sigma}}_1' = \mathbf{A}_{22}^{-1}\mathbf{a}_1$. The *sample multiple correlation coefficient* is defined to be

$$R = \sqrt{\frac{\hat{\boldsymbol{\beta}}'\hat{\mathbf{\Sigma}}_{22}\hat{\boldsymbol{\beta}}}{\hat{\sigma}_{11}}} = \sqrt{\frac{\hat{\boldsymbol{\sigma}}_1'\hat{\mathbf{\Sigma}}_{22}^{-1}\hat{\boldsymbol{\sigma}}_1}{\hat{\sigma}_{11}}} = \sqrt{\frac{\mathbf{a}_1'\mathbf{A}_{22}^{-1}\mathbf{a}_1}{a_{11}}}. \tag{4.7}$$

(R is also the maximum likelihood estimator of $\bar{R}$).

Assume that $\mathbf{x} = (X_1, \mathbf{X}_2)$ follows a p-variate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In case of $\bar{R} \neq 0$, the sampling distribution of R is complex (cf. Anderson (2003, Chapter 4)). Here we consider the case of $\bar{R} = 0$ where the sampling distribution of R is known such that

$$\frac{R^2/(p-1)}{(1-R^2)/(n-p)} \overset{\mathscr{D}}{=} F_{p-1,n-p} \,, \tag{4.8}$$

a Fisher distribution with degrees of freedom $p-1$ and $n-p$. It might be observed that $R^2/(1-R^2)$ is the statistic used to test the hypothesis that the regression of X_1 on $(X_2, \ldots, X_p)$ is 0.

As R^2 is always nonnegative so that as an estimator of $\bar{R}^2 = 0$, it has a positive bias. The *adjusted multiple correlation coefficient*

$$R^{*2} = R^2 - \frac{p-1}{n-p}(1-R^2), \tag{4.9}$$

attempts to correct this bias. Indeed, this quantity is always smaller than R^2 (unless $p = 1$ or $R^2 = 1$) and it has a smaller bias than R^2. However, R^{*2} can take negative values with positive probability, thus contradicts the original interpretation of R^2 which is a positive square. Under the classical limiting scheme where $n \to \infty$ while p is treated as a constant, both estimators are consistent, i.e. $R^2 \xrightarrow{\mathscr{P}} \bar{R}^2$ and $R^{*2} \xrightarrow{\mathscr{P}} \bar{R}^2$. The case of $\bar{R} = 0$ can be seen from (4.8): when $n \to \infty$, $F_{p-1,n-p} \xrightarrow{\mathscr{D}} \chi^2_{p-1}/(p-1)$, so that $R^2/(1-R^2) \xrightarrow{\mathscr{P}} 0$ and $R^2 \xrightarrow{\mathscr{P}} 0$.

For large-dimensional data however, we will see that these asymptotic consistencies are no longer valid. We again assume that $p/n \to y \in [0, 1)$. One might observe that if $p > n$ (or $y > 1$), the multiple correlation coefficient can still be defined but will have no reasonable estimator. For simplicity, we assume normal distributions for the observations. This restriction can be relaxed following the general theory on sample covariance matrix developed in Chapters 2 and 3.

4.3.1 Inconsistency of the sample multiple correlation coefficient

Assume that $\mathbf{x}$ has the normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then the matrix $\mathbf{A}$ has the Wishart distribution $W(N, \boldsymbol{\Sigma})$ with $N = n-1$ degrees of freedom, and thus can be written as

$$\mathbf{A} = \sum_{i=1}^{N} \mathbf{z}_i \mathbf{z}_i^* \,,$$

where the $\mathbf{z}_i$'s are i.i.d. $\mathcal{N}(0, \boldsymbol{\Sigma})$. Moreover, we represent $\mathbf{A}$ as

$$\mathbf{A} = (\mathbf{z}_1, \ldots, \mathbf{z}_N)(\mathbf{z}_1, \ldots, \mathbf{z}_N)^* = (\mathbf{y}_1, \ldots, \mathbf{y}_p)^*(\mathbf{y}_1, \ldots, \mathbf{y}_p) \,,$$

where now the $\mathbf{y}_j$'s are n-dimensional vectors. Define the matrices $\mathbf{Y}_2$ and $\mathbf{Y}_3$ such that

$$(\mathbf{y}_1, \ldots, \mathbf{y}_p) = (\mathbf{y}_1, \mathbf{Y}_2) = (\mathbf{y}_1, \mathbf{y}_2, \mathbf{Y}_3).$$

Recall the definition (4.7) of the multiple correlation coefficient,

$$R^2 = \frac{\mathbf{a}_1' \mathbf{A}_{22}^{-1} \mathbf{a}_1}{a_{11}} \,,$$

we have then

$$\begin{aligned} a_{11} &= \mathbf{y}_1'\mathbf{y}_1 \\ \mathbf{a}_1 &= \mathbf{Y}_2'\mathbf{y}_1 = \begin{pmatrix} \mathbf{y}_2'\mathbf{y}_1 \\ \mathbf{Y}_3'\mathbf{y}_1 \end{pmatrix} \\ \mathbf{A}_{22} &= \mathbf{Y}_2'\mathbf{Y}_2 \\ \mathbf{Y}_2 &= (\mathbf{y}_2, \mathbf{Y}_3) = (\mathbf{y}_2, \mathbf{y}_3, \cdots, \mathbf{y}_p). \end{aligned}$$

Since the multiple correlation coefficient R^2 is invariant with respect to lienar transformations of $\mathbf{y}_1$ or of $\mathbf{Y} - 2$, we can assume that the variables satisfy the relations

$$\begin{aligned} \mathbb{E}\mathbf{y}_j &= \mathbf{0}, & \operatorname{cov}(\mathbf{y}_j) &= \mathbf{I}_N, \\ \operatorname{cov}(\mathbf{y}_1, \mathbf{y}_2) &= \overline{R}\mathbf{I}_N, & & \\ \operatorname{cov}(\mathbf{y}_i, \mathbf{y}_j) &= \mathbf{0}, & i < j,\ (i, j) &\neq (1, 2). \end{aligned} \tag{4.10}$$

Since

$$\mathbf{A}_{22} = \begin{pmatrix} \mathbf{y}_2'\mathbf{y}_2 & \mathbf{y}_2'\mathbf{Y}_3 \\ \mathbf{Y}_3'\mathbf{y}_2 & \mathbf{Y}_3'\mathbf{Y}_3 \end{pmatrix},$$

by inversion formula for block-matrices, we have

$$\mathbf{A}_{22}^{-1} = a_{22\cdot 3}^{-1} \begin{bmatrix} 1 & -\mathbf{y}_2'\mathbf{Y}_3(\mathbf{Y}_3'\mathbf{Y}_3)^{-1} \\ -(\mathbf{Y}_3'\mathbf{Y}_3)^{-1}\mathbf{Y}_3'\mathbf{y}_2 & (\mathbf{Y}_3'\mathbf{Y}_3)^{-1} + (\mathbf{Y}_3'\mathbf{Y}_3)^{-1}\mathbf{Y}_3'\mathbf{y}_2\mathbf{y}_2'\mathbf{Y}_3(\mathbf{Y}_3'\mathbf{Y}_3)^{-1} \end{bmatrix},$$

with

$$\begin{aligned} a_{22\cdot 3} &= \mathbf{y}_2'(\mathbf{I}_N - \mathbf{Y}_3(\mathbf{Y}_3'\mathbf{Y}_3)^{-1}\mathbf{Y}_3')\mathbf{y}_2 \\ \mathbf{A}_{33\cdot 2} &= \mathbf{Y}_3'\left(\mathbf{I}_N - \frac{\mathbf{y}_2\mathbf{y}_2'}{\mathbf{y}_2'\mathbf{y}_2}\right)\mathbf{Y}_3. \end{aligned}$$

Therefore

$$R^2 = a_{11}^{-1}\left(\frac{(\mathbf{y}_1'\mathbf{y}_2 - \mathbf{y}_2'\mathbf{Y}_3(\mathbf{Y}_3'\mathbf{Y}_3)^{-1}\mathbf{Y}_3'\mathbf{y}_1)^2}{a_{22\cdot 3}} + \mathbf{y}_1'\mathbf{Y}_3(\mathbf{Y}_3'\mathbf{Y}_3)^{-1}\mathbf{Y}_3'\mathbf{y}_1\right). \tag{4.11}$$

By direct calculation and the strong law of large numbers, we have almost surely,

$$\begin{aligned} \frac{a_{11}}{n} &\to 1, \\ \frac{a_{22\cdot 3}}{n} &\to 1 - y, \\ \frac{\mathbf{y}_1'\mathbf{y}_2}{n} &\to \overline{R}, \\ \frac{1}{n}\mathbf{y}_2'\mathbf{Y}_3(\mathbf{Y}_3'\mathbf{Y}_3)^{-1}\mathbf{Y}_3'\mathbf{y}_1 &\to y\overline{R}, \\ \frac{1}{n}\mathbf{y}_1'\mathbf{Y}_3(\mathbf{Y}_3'\mathbf{Y}_3)^{-1}\mathbf{Y}_3'\mathbf{y}_1 &\to y. \end{aligned} \tag{4.12}$$

Combining (4.12) and (4.11), we find

Theorem 4.7 *For Gaussian observations and assume that $p/n \to y \in [0, 1)$,*

$$R^2 \xrightarrow{\text{a.s.}} (1 - y)\overline{R}^2 + y. \tag{4.13}$$

Therefore, under the $p \propto n$ scheme ($y > 0$) the sample multiple correlation coefficient will almost surely over-estimate the population multiple correlation coefficient unless $\overline{R} = 1$ (an useless situation). Another striking consequence of the theorem is that the adjusted sample multiple correlation coefficient R^{*2} remains consistent under these large-dimensional schemes, as it can be checked using (4.9). Even when possible negative values of R^{*2} are rounded to 0 (i.e. considering $R^{*2}I\{R^{*2} \geq 0\}$, the modified estimator remains consistent,

Finally, notice that the scheme $p \ll n$ ($y = 0$) extends the classical limit scheme where the sample multiple correlation coefficient is consistent.

4.3.2 CLT for the sample multiple correlation coefficient

In this section we will find a central limit theorem for R^2 under the large-dimensional scheme. For a given $\mathbf{Y}_3$, we may find a $N \times (p-2)$ random matrix $\mathbf{E}$ satisfying

$$\mathbf{E}\mathbf{E}' = \mathbf{Y}_3(\mathbf{Y}_3'\mathbf{Y}_3)^{-1}\mathbf{Y}_3', \quad \mathbf{E}'\mathbf{E} = \mathbf{I}_{p-2}. \tag{4.14}$$

Moreover, we can find another $N \times (N-p+2)$ random matrix $\mathbf{F}$ such that $\mathbf{Q} = (\mathbf{E}, \mathbf{F})$ is a N-dimensional orthogonal matrix. For $j = 1, 2$, define

$$\mathbf{u}_j = \mathbf{Q}'\mathbf{y}_j = \begin{pmatrix} \mathbf{v}_{1j} \\ \mathbf{v}_{2j} \end{pmatrix},$$

where $\mathbf{v}_{1j}$ has dimension $(p-2)$ and $\mathbf{v}_{2j}$ has dimension $(N-p+2)$. It is easy to see that $(\mathbf{u}_1, \mathbf{u}_2)$ is a Gaussian vector with mean 0 whose covariance matrix satisfy

$$\begin{aligned} \operatorname{cov}(\mathbf{u}_j) &= \mathbf{I}_N, \qquad \operatorname{cov}(\mathbf{u}_1, \mathbf{u}_2) = \mathbf{0}, \\ \operatorname{cov}(\mathbf{v}_{11}, \mathbf{v}_{21}) &= \overline{R}\,\mathbf{I}_{p-2}, \qquad \operatorname{cov}(\mathbf{v}_{12}, \mathbf{v}_{22}) = \overline{R}\mathbf{I}_{N-p+2}, \\ \operatorname{cov}(\mathbf{v}_{j1}, \mathbf{v}_{j2}) &= \mathbf{0}, \qquad j = 1, 2. \end{aligned} \tag{4.15}$$

Since the distribution of $(\mathbf{u}_1, \mathbf{u}_2)$ is independent of $\mathbf{Y}_3$, they are independent. In correspondence with Eq.(4.11), we have

$$\begin{aligned} a_{11} &\overset{\mathscr{D}}{=} \mathbf{u}_1'\mathbf{u}_1 = \mathbf{v}_{11}'\mathbf{v}_{11} + \mathbf{v}_{12}'\mathbf{v}_{12} \\ \mathbf{y}_1'\mathbf{y}_2 - \mathbf{y}_2'\mathbf{Y}_3(\mathbf{Y}_3'\mathbf{Y}_3)^{-1}\mathbf{Y}_3'\mathbf{y}_1 &\overset{\mathscr{D}}{=} \mathbf{v}_{22}'\mathbf{v}_{12} \\ \mathbf{y}_1'\mathbf{Y}_3(\mathbf{Y}_3'\mathbf{Y}_3)^{-1}\mathbf{Y}_3'\mathbf{y}_1 &\overset{\mathscr{D}}{=} \mathbf{v}_{11}'\mathbf{v}_{11} \\ a_{22\cdot 3} &\overset{\mathscr{D}}{=} \mathbf{v}_{22}'\mathbf{v}_{22}. \end{aligned}$$

By standard central limit theorem,

$$\begin{aligned} \frac{1}{\sqrt{N}}(\mathbf{v}_{11}'\mathbf{v}_{11} - (p-2)) &\xrightarrow{\mathscr{D}} W_1 \sim \mathcal{N}(0, 2y) \\ \frac{1}{\sqrt{N}}(\mathbf{v}_{12}'\mathbf{v}_{12} - (N-p+2)) &\xrightarrow{\mathscr{D}} W_2 \sim \mathcal{N}(0, 2(1-y)) \\ \frac{1}{\sqrt{N}}(\mathbf{v}_{22}'\mathbf{v}_{22} - (N-p+2)) &\xrightarrow{\mathscr{D}} W_3 \sim \mathcal{N}(0, 2(1-y)) \\ \frac{1}{\sqrt{N}}(\mathbf{v}_{22}'\mathbf{v}_{12} - (N-p+2)\overline{R}) &\xrightarrow{\mathscr{D}} W_4 \sim \mathcal{N}\Big(0, (1+\overline{R}^2)(1-y)\Big), \end{aligned} \tag{4.16}$$

where the random variables $\{W_j\}_{1\le j\le 4}$ are independent. Plugging these relationships (4.16) into Eq.(4.11) and by the multivariate Delta method, we obtain the following theorem.

Theorem 4.8 *Assume $p \to \infty$, $n \to \infty$ and $p/y \to y \in (0,1)$. Then*

$$\sqrt{N}\left[R^2 - (1-y_n)\overline{R}^2 - y_n\right] \xrightarrow{\mathscr{D}} \mathcal{N}(0, \sigma^2(y)), \tag{4.17}$$

where $y_n = (p-2)/N$ and

$$\sigma^2(y) = 2(1-y)[y + 2\overline{R}^2 + (4-y)\overline{R}^4].$$

Proof By expanding the right-hand side of (4.11) by Taylor's formula and using (4.16), we have

$$\begin{aligned}
&\sqrt{N}\left\{R^2 - (1-y_n)\overline{R}^2 - y_n\right\}\\
&\simeq -\left[(1-y_n)\overline{R}^2 + y_n\right](W_1 + W_2) + \frac{2(1-y_n)\overline{R}}{1-y_n}W_4 - \frac{(1-y_n)^2\overline{R}^2}{(1-y_n)^2}W_3 + W_1\\
&\to (1-y)(1-\overline{R}^2)W_1 - [(1-y)\overline{R}^2 + y]W_2 + 2\overline{R}W_4 - \overline{R}^2 W_3.
\end{aligned}$$

The asymptotic variance follows from the last formula and the proof is complete. □

Applying the theorem and using the delta method, we obtain the following results.

Corollary 4.9 *Under the large-dimensional scheme, we have*

$$\sqrt{N}\left(\frac{R^2}{1-R^2} - \frac{(1-y_n)\overline{R}^2 + y_n}{(1-y_n)(1-\overline{R}^2)}\right) \xrightarrow{\mathscr{D}} \mathcal{N}(0, \sigma_f^2(y)),$$

where

$$\sigma_f^2(y) = \frac{2[y + 2\overline{R}^2 + (4-y)\overline{R}^4]}{(1-y)^3(1-\overline{R}^2)^4}.$$

Corollary 4.10 *Under the large-dimensional scheme, we have*

$$\sqrt{N}\left(R - \sqrt{(1-y_n)\overline{R}^2 + y_n}\right) \xrightarrow{\mathscr{D}} \mathcal{N}(0, \sigma_o^2(y)),$$

where

$$\sigma_o^2(y) = \frac{(1-y)[y + 2\overline{R}^2 + (4-y)\overline{R}^4]}{2[(1-y)\overline{R}^2 + y]}.$$

One might notice that although the above results are developed using the large-dimensional theory, they remain valid even for small data dimension p. Indeed firstly, $y_n = (p-2)/N$ is always a positive number; and secondly, the derivations of Theorem 4.8 and Corollary 4.9 and 4.10 are all valid if $y = 0$.

Finally, hypothesis testing and confidence intervals can be constructed using Corollary 4.9 or 4.10 and they are left to the reader.

Notes

Complementary results to the analysis of the multiple correlation coefficient in Section 4.3 can be found in Zheng et al. (2014).

5

Testing hypotheses of equality of covariance matrices

5.1 Introduction

In this chapter the problem of testing hypotheses on covariance matrices is considered. At some stage, hypotheses on means of populations are also added into consideration. Traditionally, two situations are distinguished: the first concerns a single population for testing the hypothesis that a covariance matrix is equal to a given matrix, or having a specific structure, i.e. diagonal, proportional to the identity matrix. The second situation concerns two or more populations where a typical hypothesis is that these populations have a same covariance matrix.

We start by a review of traditional multivariate procedures for these tests. Most of these material can be found in more details in Anderson (2003, Chapter 10). Then we develop corrections or adjustments of these procedures to cope with large-dimensional data.

5.2 Testing equality between several covariance matrices

Let $\mathcal{N}(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$, $1 \leq g \leq q$ be q normal distributions of dimension p. For each of these distributions, say $\mathcal{N}(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$, we collect a sample $\mathbf{x}_{gk}$, $k = 1, \cdots, n_g$, of size n_g. Let $n = n_1 + \cdots + n_q$ be the total sample size. The aim is to test the hypothesis that the q populations have a same covariance matrix, that is

$$H_{v0} : \boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_q. \tag{5.1}$$

Let Ω be the general parameter space with arbitrary positive definite matrices $\boldsymbol{\Sigma}_g$ and mean vectors $\boldsymbol{\mu}_g$. Let also $\omega \subset \Omega$ be the subset restricted by the null hypothesis $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_q$.

Consider first the likelihood ratio test for H_{v0}. The following results are well-known from classical multivariate analysis for normal populations (Anderson, 2003, §10.2). First, the maximum likelihood estimators of $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ in Ω are

$$\widehat{\boldsymbol{\mu}}_{g\Omega} = \bar{\mathbf{x}}_g, \quad \widehat{\boldsymbol{\Sigma}}_{g\Omega} = \frac{1}{n_g}\mathbf{A}_g, \qquad g = 1, \cdots, q, \tag{5.2}$$

where

$$\bar{\mathbf{x}}_g = \frac{1}{n_g}\sum_{k=1}^{n_g} \mathbf{x}_{gk}, \quad \mathbf{A}_g = \sum_{k=1}^{n_g} (\mathbf{x}_{gk} - \bar{\mathbf{x}}_g)(\mathbf{x}_{gk} - \bar{\mathbf{x}}_g)'. \tag{5.3}$$

The corresponding maximum likelihood is

$$\mathscr{L}_\Omega = (2\pi e)^{-\frac{1}{2}pn} \prod_{g=1}^{q} \left|\mathbf{A}_g/n_g\right|^{-\frac{1}{2}n_g} . \tag{5.4}$$

Under the null hypothesis, the maximum likelihhod estimators are unchanged for the mean vectors $\boldsymbol{\mu}_g$. The maximum likelihood estimator for the common covariance matrix $\boldsymbol{\Sigma}$ is now

$$\widehat{\boldsymbol{\Sigma}}_\omega = \frac{1}{n}\mathbf{A}, \quad \mathbf{A} = \sum_{g=1}^{q} \mathbf{A}_g.$$

The corresponding maximum likelihood is

$$\mathscr{L}_\omega = (2\pi e)^{-\frac{1}{2}pn} \left|\mathbf{A}/n\right|^{-\frac{1}{2}n} . \tag{5.5}$$

Therefore, the likelihood ratio for testing (5.1) is

$$\lambda_1 = \frac{\mathscr{L}_\omega}{\mathscr{L}_\Omega} = \frac{\prod_{g=1}^{q} \left|\mathbf{A}_g/n_g\right|^{-\frac{1}{2}n_g}}{\left|\mathbf{A}/n\right|^{\frac{1}{2}n}} . \tag{5.6}$$

The equality hypothesis is rejected at level α if

$$\lambda_1 \leq C_1(\alpha), \tag{5.7}$$

for some critival value $C_1(\alpha)$.

To improve the accuracy of the likelihood ratio λ_1, Bartlett (1937) suggested substituting the corresponding degrees of freedom for the sample sizes in the $\mathbf{A}_g$'s. This leads to consider the test statistic

$$V_1 = \frac{\prod_{g=1}^{q} \left|\mathbf{A}_g\right|^{\frac{1}{2}N_g}}{\left|\mathbf{A}\right|^{\frac{1}{2}N}}, \tag{5.8}$$

where

$$N_g = n_g - 1, \; 1 \leq g \leq q, \quad N = N_1 + \cdots + N_q = n - q.$$

A first insignt about this corrected statistic can ba gained by considering the particular case of two univariate normal populations, that is, $p = 1$ and $q = 2$. The statistic (5.8) becomes

$$V_1 = \frac{N_1{}^{\frac{1}{2}N_1} N_2{}^{\frac{1}{2}N_2} (s_1^2)^{\frac{1}{2}N_1} (s_2^2)^{\frac{1}{2}N_2}}{(N_1 s_1^2 + N_2 s_2^2)^{\frac{1}{2}N}} = \frac{N_1{}^{\frac{1}{2}N_1} N_2{}^{\frac{1}{2}N_2} F^{\frac{1}{2}N_1}}{\{N_1 F + N_2\}^{\frac{1}{2}N}}, \tag{5.9}$$

where s_1^2 and s_2^2 are the usual unbiased estimators of the two population variances σ_1^2 and σ_2^2, respectively. Here, $F = s_1^2/s_2^2$ the classical F-ratio with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. The likelihood ratio test thus reduces to the well-known Fisher test using this F-ratio.

It is well-known that this likelihood ratio test is invariant with respect to changes of location within populations and a common linear transformation. An alternative invariant test procedure (Nagao, 1973b) is based on the criterion

$$\frac{1}{2}\sum_{g=1}^{q}(n_g - k_g)\mathrm{tr}(\mathbf{S}_g\mathbf{S}^{-1} - \mathbf{I})^2 = \frac{1}{2}\sum_{g=1}^{q}(n_g - k_g)\mathrm{tr}(\mathbf{S}_g - \mathbf{S})\mathbf{S}^{-1}(\mathbf{S}_g - \mathbf{S})\mathbf{S}^{-1}, \tag{5.10}$$

where $\mathbf{S}_g = \mathbf{A}_g/N_g$ and $\mathbf{S} = \mathbf{A}/N$. Here $k_g = (n_g - 1)/(n - q)$ and $\sum k_g = 1$.

5.3 Testing equality of several multivariate normal distributions

In this section we want to test the full identity between q multivariate normal distributions. Therefore, in addition to equality of the population covariance matrices in (5.1), we also request equality of the population means. The hypothesis to be tested is

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_q, \quad \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots, = \boldsymbol{\Sigma}_q. \tag{5.11}$$

As in the previous section, let $\mathbf{x}_{gk}, k = 1, \cdots, n_g$, be an observation from $\mathcal{N}(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$, $g = 1, \cdots, q$. Then Ω is the general parameter space of $\{\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\}, g = 1, \cdots, q$, and ω' consists of the sub-space of parameters restricted by (5.11).

The maximum of the likelihood function on Ω is given by Eq. (5.4). Under H_0, the maximum likelihood estimators of the common mean and covariance matrix are

$$\hat{\boldsymbol{\mu}}_{\omega'} = \bar{\mathbf{x}} = \frac{1}{n}\sum_{g,k} \mathbf{x}_{gk}\ , \qquad \hat{\boldsymbol{\Sigma}}_{\omega'} = \frac{1}{n}\mathbf{B} \tag{5.12}$$

where

$$\begin{aligned}\mathbf{B} &= \sum_{g=1}^{q}\sum_{k=1}^{n_g}(\mathbf{x}_{gk} - \bar{\mathbf{x}})(\mathbf{x}_{gk} - \bar{\mathbf{x}})' \\ &= \mathbf{A} + \sum_{g=1}^{q} n_g(\bar{\mathbf{x}}_g - \bar{\mathbf{x}})(\bar{\mathbf{x}}_g - \bar{\mathbf{x}})'.\end{aligned}$$

The maximum of the likelihood function on ω' is

$$\mathscr{L}_{\omega'} = (2\pi e)^{-\frac{1}{2}pn}\, |\mathbf{B}/n|^{-\frac{1}{2}n}\ . \tag{5.13}$$

Therefore, the likelihood ratio criterion for the hypothesis H_0 is

$$\lambda = \frac{\mathscr{L}_{\omega'}}{\mathscr{L}_{\Omega}} = \frac{\prod_{g=1}^{q} |\mathbf{A}_g/n_g|^{\frac{1}{2}n_g}}{|\mathbf{B}/n|^{\frac{1}{2}n}}\ . \tag{5.14}$$

Notice that since

$$\lambda = \frac{\mathscr{L}_{\omega'}}{\mathscr{L}_{\omega}} \cdot \frac{\mathscr{L}_{\omega}}{\mathscr{L}_{\Omega}},$$

the criterion λ is the product of the likelihood ratio criterion λ_1 in (5.6) and the criterion, say λ_m in §**??** for testing the hypothesis that the means are equal.

Let

$$V_2 = \frac{|\mathbf{A}|^{\frac{1}{2}N}}{|\mathbf{B}|^{\frac{1}{2}N}} = \lambda_m^{N/n};$$

this is equivalent to λ_m for testing on the means. We might consider as in Bartlett's correction,

$$V = V_1V_2 = \frac{\prod_{g=1}^{q} |\mathbf{A}_g|^{\frac{1}{2}N_g}}{|\mathbf{B}|^{\frac{1}{2}N}}. \tag{5.15}$$

In order to derive the null distribution of the statistic V, let us first consider V_1 given by (5.8). Define

$$V_{1g} = \frac{|\mathbf{A}_1 + \cdots + \mathbf{A}_{g-1}|^{\frac{1}{2}(N_1+\cdots+N_{g-1})}|\mathbf{A}_g|^{\frac{1}{2}N_g}}{|\mathbf{A}_1 + \cdots + \mathbf{A}_g|^{\frac{1}{2}(N_1+\cdots+N_g)}}, \quad g = 2, \cdots, q, \tag{5.16}$$

then

$$V_1 = \prod_{g=2}^{q} V_{1g}.$$

Theorem 5.1 $V_{12}, \ldots, V_{1q}$ *defined by (5.16) are independent when* $\mathbf{\Sigma}_1 = \cdots = \mathbf{\Sigma}_q$*,* $n_g > p$*,* $g = 1, \cdots, q$*.*

The proofs of this theorem and of the following characterisation can be found in Anderson (2003, §10.3).

Theorem 5.2

$$V_1 = \prod_{g=2}^{q} \left\{ \prod_{i=1}^{p} X_{ig}^{\frac{1}{2}(N_1+\cdots+N_{g-1})}(1 - X_{ig})^{\frac{1}{2}N_g} \cdot \prod_{i=2}^{p} Y_{ig}^{\frac{1}{2}(n_1+\cdots+n_g-g)} \right\},$$

where the X's and Y's are independent, X_{ig} *has the* $\beta\left[\frac{1}{2}(n_1 + \cdots + n_{g-1} - g - i + 2), \frac{1}{2}(n_g - i)\right]$ *distribution, and* Y_{ig} *has the* $\beta\left[\frac{1}{2}(n_1 + \cdots + n_g - g) - i + 1, \frac{1}{2}(i - 1)\right]$ *distribution.*

Now consider the likelihood ratio criterion λ given in (5.14) for testing the identity hypothesis (5.11) This is equivalent to the criterion

$$W = \frac{\prod_{g=1}^{q} |\mathbf{A}_g|^{\frac{1}{2}n_g}}{|\mathbf{A}_1 + \cdots + \mathbf{A}_q|^{\frac{1}{2}(n_1+\cdots+n_g)}} \cdot \frac{|\mathbf{A}_1 + \cdots + \mathbf{A}_q|^{\frac{n}{2}}}{|\mathbf{A}_1 + \cdots + \mathbf{A}_q + \sum_{g=1}^{q} n_g(\bar{\mathbf{x}}_g - \bar{\mathbf{x}})(\bar{\mathbf{x}}_g - \bar{\mathbf{x}})'|^{\frac{n}{2}}}. \tag{5.17}$$

The preceding two factors are independent because the first factor is independent of $\mathbf{A}_1 + \cdots + \mathbf{A}_q$ and of $\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_q$.

The following theorem can be found in Anderson (2003, §10.4).

Theorem 5.3

$$\mathbf{W} = \prod_{g=2}^{q} \left\{ \prod_{i=1}^{p} X_{ig}^{\frac{1}{2}(n_1+\cdots+n_{g-1})}(1 - X_{ig})^{\frac{1}{2}n_g} \cdot \prod_{i=2}^{p} Y_{ig}^{\frac{1}{2}(n_1+\cdots+n_g)} \right\} \prod_{i=1}^{p} Z_i^{\frac{1}{2}n},$$

where the X's, Y's, and Z's are independent, X_{ig} *has the* $\beta\left[\frac{1}{2}(n_1 + \cdots + n_{g-1} - g - i + 2), \frac{1}{2}(n_g - i)\right]$ *distribution,* Y_{ig} *has the* $\beta\left[\frac{1}{2}(n_1 + \cdots + n_g - g) - i + 1, \frac{1}{2}(i - 1)\right]$ *distribution, and* Z_i *has the* $\beta\left[\frac{1}{2}n - i, \frac{1}{2}(q - 1)\right]$ *distribution.*

Unfortunately, the exact distributions given in Theorems 5.1, 5.2 and 5.3 are very complex, especially for large values of p or q. It is therefore reasonable to seek some asymptotic approximation for these distributions.

Recall that $n_g - 1 = k_g(n - q)$, where $\sum_{g=1}^{q} k_g = 1$. The expansions are in terms of $n - q$ increasing with $k_1, \cdots, k_q$ fixed (we could assume only $\lim N_g/(n - q) = k_g > 0$). Let also $\varphi_m(z) = \mathbb{P}(\chi_m^2 \leq z)$.

(i) For

$$\lambda_1^* = V_1 \cdot \frac{n^{\frac{1}{2}pN}}{\prod_{g=1}^q n_g^{\frac{1}{2}pN_g}} = V_1 \cdot \prod_{g=1}^{q} \left(\frac{n-q}{n_g - 1} \right)^{\frac{1}{2}pN_g}$$
$$= \left[\prod_{g=1}^{q} \left(\frac{1}{k_g} \right)^{k_g} \right]^{\frac{1}{2}pN} V_1,$$

and with

$$\rho = 1 - \left(\sum_{g=1}^{q} \frac{1}{n_g} - \frac{1}{n} \right) \frac{2p^2 + 3p - 1}{6(p+1)(q-1)},$$
$$\omega_2 = \frac{p(p+1)}{48\rho^2} \left[(p-1)(p+2) \left(\sum_{g=1}^{q} \frac{1}{{N_g}^2} - \frac{1}{N^2} \right) - 6(q-1)(1-\rho)^2 \right],$$

we have

$$\mathbb{P}\{-2\rho \log \lambda_1^* \le z\} = \varphi_f(z) + \omega_2 \left[\varphi_{f+4}(z) - \varphi_f(z) \right] + O(n^{-3}). \tag{5.18}$$

(ii) For $\lambda = Wn^{\frac{1}{2}pn} \prod_{g=1}^q n_g^{-\frac{1}{2}pn_g}$, with

$$\rho = 1 - \left(\sum_{g=1}^{q} \frac{1}{n_g} - \frac{1}{n} \right) \frac{2p^2 + 9p + 11}{6(p+3)(q-1)},$$

and

$$\omega_2 = \frac{p(p+3)}{48\rho^2} \left[\sum_{g=1}^{q} \left(\frac{1}{n_g^2} - \frac{1}{n^2} \right) (p+1)(p+2) - 6(1-\rho)^2(q-1) \right],$$

we have

$$\Pr\{-2\rho \log \lambda \le z\} = \varphi_f(z) + \omega_2 \left[\varphi_{f+4}(z) - \varphi_f(z) \right] + O(n^{-3}). \tag{5.19}$$

5.4 The sphericity test

Consider a p-dimensional normal population $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The population is *spherical* if $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$ for some unspecified positive constant σ^2. This is probably the simplest structure one may assume about a covariance matrix. To test this sphericity hypothesis, namely, the hypothesis

$$H_0: \quad \boldsymbol{\Sigma} = \sigma^2 \mathbf{I}, \quad \sigma^2 > 0, \tag{5.20}$$

consider a sample $\mathbf{x}_1, \cdots, \mathbf{x}_n$ from the population $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let Ω be the general parameter space with arbitrary pair $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\omega \subset \Omega$ be the subset restricted by the sphericity hypothesis H_0.

An important characterisation of a spherical covariance matrix is the following. Let $\phi_1, \cdots, \phi_p$ be the (nonnegative) eigenvalues of $\boldsymbol{\Sigma}$. By convexity, the ratio between their geometric and arithmetic means is not larger than one, that is,

$$\frac{\prod_{i=1}^p \phi_i^{1/p}}{\sum_{i=1}^p \phi_i / p} = \frac{|\boldsymbol{\Sigma}|^{1/p}}{\mathrm{tr}\boldsymbol{\Sigma}/p} \le 1. \tag{5.21}$$

The equality holds if and only if the ϕ_is are equal, that is, $\mathbf{\Sigma}$ is spherical.

Consider the likelihood ratio test which has a simple form here. First, the maximum likelihood estimators of $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ in Ω are

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k, \quad \hat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{A}, \tag{5.22}$$

where

$$\mathbf{A} = \sum_{k=1}^{n}(\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})'.$$

The corresponding maximum likelihood is

$$\mathscr{L}_\Omega = (2\pi e)^{-\frac{1}{2}pn}\left|\hat{\mathbf{\Sigma}}\right|^{-\frac{1}{2}n} . \tag{5.23}$$

Under the null hypothesis, the maximum likelihhod estimator for the mean is still $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$. Let $\bar{\mathbf{x}} = (\bar{x}_1, \ldots, \bar{x}_p)'$, and $\mathbf{x}_k = (\bar{x}_{k1}, \ldots, \bar{x}_{kp})'$, The maximum likelihood estimator for the covariance matrix $\mathbf{\Sigma}$ is now

$$\hat{\mathbf{\Sigma}}_\omega = \hat{\sigma}^2\mathbf{I},$$

where

$$\hat{\sigma}^2 = \frac{1}{np}\sum_{i=1}^{p}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)^2 = \frac{1}{np}\sum_{k=1}^{n}||\bar{\mathbf{x}}_k - \bar{\mathbf{x}}||^2 = \frac{1}{np}\operatorname{tr}\mathbf{A}.$$

The corresponding maximum likelihood is

$$\mathscr{L}_\omega = (2\pi e)^{-\frac{1}{2}pn}|\hat{\sigma}^2\mathbf{I}|^{-\frac{1}{2}n}. \tag{5.24}$$

Therefore, the likelihood ratio for testing (5.20) is

$$\lambda = \frac{\mathscr{L}_\omega}{\mathscr{L}_\Omega} = \frac{|\widehat{\mathbf{\Sigma}}|^{\frac{1}{2}n}}{|\hat{\sigma}^2\mathbf{I}|^{\frac{1}{2}n}} = \frac{|\mathbf{A}|^{\frac{1}{2}n}}{(\operatorname{tr}\mathbf{A}/p)^{\frac{1}{2}pn}}. \tag{5.25}$$

The sphericity hypothesis is rejected at level α if

$$\lambda \le C(\alpha),$$

for some critival value $C(\alpha)$.

Let $l_1, \cdots, l_p$ be the eigenvalues of the sample covariance matrix $\mathbf{S} = (n-1)^{-1}\mathbf{A}$. Then

$$\lambda = \left(\frac{\prod l_i^{1/p}}{\sum l_i/p}\right)^{\frac{1}{2}pn}, \tag{5.26}$$

that is, a power of the ratio between their geometric and arithmetic means. This reminds the characterisation of sphericity (5.21).

Again, the exact distribution of the likelihood ratio λ under the null hypothesis is very complicate. Similar to previous asymptotic expansions, Bartlett-type expansion for the distribution of $\log\lambda$ exists. Let $f = \frac{1}{2}p(p+1) - 1$ and

$$\rho = 1 - \frac{2p^2 + p + 2}{6p(n-1)},$$

$$\omega_2 = \frac{(p+2)(p-1)(p-2)(2p^3 + 6p^2 + 3p + 2)}{288p^2(n-1)^2\rho^2},$$

we have, with $\varphi_m(z) = \mathbb{P}(\chi^2_m \le z)$,

$$\mathbb{P}\{-2\rho \log \lambda \le z\} = \varphi_f(z) + \omega_2 \left\{\varphi_{f+4}(z) - \varphi_f(z)\right\} + O(n^{-3}). \tag{5.27}$$

Using modern software, we can compute factors $c(n, p, \alpha)$ satisfying

$$\mathbb{P}\{-2\rho \log \lambda \le c(n, p, \alpha)\chi^2_f(\alpha)\} = \alpha,$$

or calculate directly p-values. Here $\chi^2_f(\alpha)$ denotes the αth upper quantile of χ^2_f, i.e $\varphi_f^{-1}(\alpha)$.

Other interesting tests exist for the sphericity hypothesis (Anderson, 2003, §10.7.5). The null hypothesis $H : \mathbf{\Sigma} = \sigma^2\mathbf{I}$ is invariant with respect to transformations $\mathbf{X}^* = c\mathbf{QX} + \boldsymbol{\nu}$, where c is a scalar and $\mathbf{Q}$ is an orthogonal matrix. It can be shown that the maximum invariants with respect to these transformations are the $p-1$ ratios $l_1/l_2, \ldots, l_{p-1}/l_p$, where the l_is are the eigenvalues of the sample covariance matrix $\mathbf{S}$. Any invariant test is based on functions of these ratios. The likelihood ratio λ is an invariant test. Another invariant test is proposed by John (1971) with the statistic

$$\begin{aligned} &\frac{1}{2}n\mathrm{tr}\left(\mathbf{S} - \frac{\mathrm{tr}\mathbf{S}}{p}\mathbf{I}\right)\frac{p}{\mathrm{tr}\mathbf{S}}\left(\mathbf{S} - \frac{\mathrm{tr}\mathbf{S}}{p}\mathbf{I}\right)\frac{p}{\mathrm{tr}\mathbf{S}} \\ &= \frac{1}{2}n\mathrm{tr}\left(\frac{p}{\mathrm{tr}\mathbf{S}}\mathbf{S} - \mathbf{I}\right)^2 = \frac{1}{2}n\left[\frac{p^2}{(\mathrm{tr}\mathbf{S})^2}\mathrm{tr}\mathbf{S}^2 - p\right] \\ &= \frac{1}{2}n\left[\frac{p^2}{(\sum_{i=1}^p l_i)^2}\sum_{i=1}^p l_i^2 - p\right] = \frac{1}{2}n\frac{\sum_{i=1}^p (l_i - \bar{l})^2}{\bar{l}^2}, \end{aligned} \tag{5.28}$$

where $\bar{l} = \sum_{i=1}^p l_i/p$. This equation connects a squared loss function for the sample covariance matrix $\mathbf{S}$ (left-hand side) with the coefficient of variation of its eigenvalues (right-hand side). It is worth mentioning yet another interesting invariant test which uses the statistic l_1/l_p, that is, the conditioning number of the matrix.

5.5 Testing equality about one covariance matrix

Consider again a p-dimensional normal population $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$. The aim here is to test whether $\mathbf{\Sigma}$ is equal to a given matrix, say $\mathbf{\Sigma}_0$. Because the population is normal and $\mathbf{\Sigma}_0$ is given, the problem is reduced to test the equality to the identity matrix if we transform data, say $\mathbf{x}$, to $\mathbf{\Sigma}_0^{-1/2}\mathbf{x}$. Therefore, without loss of generality, we consider testing the identity hypothesis

$$H_0 : \ \mathbf{\Sigma} = \mathbf{I}. \tag{5.29}$$

Consider a sample $\mathbf{x}_1, \cdots, \mathbf{x}_n$ from the population $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$. Similar to the calculations for the sphericity test in §5.4, the maximum likelihood under the alternative hypothesis is

$$\mathscr{L}_\Omega = (2\pi e)^{-\frac{1}{2}pn}\left|\hat{\mathbf{\Sigma}}\right|^{-\frac{1}{2}n}, \tag{5.30}$$

where

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{A}, \quad \mathbf{A} = \sum_{k=1}^n (\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})', \quad \bar{\mathbf{x}} = \frac{1}{n}\sum_{k=1}^n \mathbf{x}_k.$$

The maximum likelihood under the null hypothesis is

$$\mathscr{L}_\omega = (2\pi)^{-\frac{1}{2}pn} \exp\left[-\frac{1}{2}\sum_{k=1}^{n}(\mathbf{x}_k - \bar{\mathbf{x}})'(\mathbf{x}_k - \bar{\mathbf{x}})\right] = (2\pi)^{-\frac{1}{2}pn} \exp\left[-\frac{1}{2}\mathrm{tr}\mathbf{A}\right]. \tag{5.31}$$

Therefore, the likelihood ratio for testing (5.29) is

$$\lambda_1 = \frac{(2\pi)^{-\frac{1}{2}pn} \exp\left[-\frac{1}{2}\mathrm{tr}\mathbf{A}\right]}{(2\pi e)^{-\frac{1}{2}pn}|\mathbf{A}/n|^{-\frac{1}{2}n}} = \left(\frac{e}{n}\right)^{\frac{1}{2}pn} |\mathbf{A}|^{\frac{1}{2}n} e^{-\frac{1}{2}\mathrm{tr}\mathbf{A}}. \tag{5.32}$$

Sugiura and Nagao (1968) recommended a slightly modified version

$$\lambda_1^* = \left(\frac{e}{N}\right)^{\frac{1}{2}pN} |\mathbf{A}|^{\frac{1}{2}N} e^{-\frac{1}{2}\mathrm{tr}\mathbf{A}} = e^{\frac{1}{2}pN}\left(|\mathbf{S}|e^{-\mathrm{tr}\mathbf{S}}\right)^{\frac{1}{2}N}, \tag{5.33}$$

where $N = n - 1$ and $\mathbf{S} = \mathbf{A}/N$, is unbiased. Note that

$$-\frac{2}{N}\log\lambda_1^* = \mathrm{tr}\mathbf{S} - \log|\mathbf{S}| - p = L_1(\mathbf{I}, \mathbf{S}), \tag{5.34}$$

where $L_1(\mathbf{I}, \mathbf{S})$ is a likelihood-based loss function for estimating $\mathbf{I}$ by $\mathbf{S}$. This is also easily expressed using the eigenvalues l_is of $\mathbf{S}$, that is,

$$L_1(\mathbf{I}, \mathbf{S}) = \sum_{i=1}^{p}(l_i - \log l_i - 1). \tag{5.35}$$

Naturally, the distance vanishes when $\mathbf{S} = \mathbf{I}$, that is, $l_i = 1$ for all i.

The distribution of the modified likelihood ratio criterion has also an asymptotic expansionof Bartlett-type: with $\varphi_m(z) = \mathbb{P}(\chi_m^2 \le z)$,

$$\mathbb{P}\{-2\log\lambda_1^* \le z\} = \varphi_f(z) + \frac{\gamma_2}{\rho^2(n-1)^2}\left\{\varphi_{f+4}(z) - \varphi_f(z)\right\} + O(n^{-3}), \tag{5.36}$$

where $f = \frac{1}{2}p(p+1)$ and

$$\rho = 1 - \frac{2p^2 + 3p - 1}{6n(p+1)},$$
$$\gamma_2 = \frac{p(2p^4 + 6p^3 + p^2 - 12p - 13)}{288(p+1)}.$$

5.6 Testing hypotheses of equality of large covariance matrices

5.6.1 Correction of the likelihood ratio test for equality: one-matrix case

Testing whether a population covariance matrix equals a given matrix for a normal population can be reduced to the problem of testing whether a population covariance matrix is unit matrix. In section 5.5, we derive that the criterion of likelihood ratio for the test is as given in (5.34), i.e.

$$\mathrm{LRT}_1 = \mathrm{tr}\mathbf{S} - \log|\mathbf{S}| - p.$$

In section 5.5, we discuss the precise distribution of the criterion and derive its asymptotic expansion. However, as mentioned before, this precise distribution is not easy to use, and the asymptotic expansion has large error under large dimensional structure. Therefore, new criteria are needed to deal with large-dimensional data.

Theorem 5.4 *Assume that $p \wedge n \to \infty$ and $p/n \to y \in (0,1)$. Then*

$$LRT_1 - pd_1(y_N) \xrightarrow{\mathscr{D}} \mathcal{N}(\mu_1, \sigma_1^2),$$

where $N = n - 1$, $y_N = p/N$ and

$$d_1(y) = 1 + \frac{1-y}{y}\log(1-y),$$
$$\mu_1 = -\frac{1}{2}\log(1-y),$$
$$\sigma_1^2 = -2\log(1-y) - 2y.$$

Proof The proof is a simple application of the CLT 3.4 and the substitution principle in Theorem 3.16 to the linear spectral statistic

$$\frac{1}{p}\mathrm{LRT}_1 = F^{\mathbf{S}}(g), \quad g(x) = x - \log x - 1\ ,$$

where $F^{\mathbf{S}}$ denotes the ESD of the unbiased sample covariance matrix $\mathbf{S}$. In particular, the value of the centring parameter $d_1(y_N)$ is calculated in Example 2.11. The limiting parameters μ_1 and σ_1^2 can be derived using Proposition 3.8. □

Notice that as shown in its proof, this theorem is also valid for a non-normal population. In this case, the limiting parameters μ_1 and σ_1^2 are to be adapted according to Proposition 3.8.

Using the theorem, a test can be designed with critical region

$$\mathrm{LRT}_1 \geq pd_1(y_N) + \mu_1 + \sigma_1 z_\alpha\ ,$$

where z_α is the αth upper quantile of standard normal. This test is called the *corrected likelihood ratio test* for the equality hypothesis.

Using simulation, this criterion is compared to the traditional likelihood ratio λ^* in (5.34) with limiting distribution given in (5.36). Empirical results based on 10,000 independent replications are reported in Table 5.1.

Table 5.1 *Comparison of Type I error and power between corrected likelihood ratio test and traditional likelihood ratio test*

	Corrected likelihood ratio test			Traditional likelihood ratio test(Wilks)	
(p,n)	Type I error	5% difference	Power	Type I error	Power
(5, 500)	0.0803	0.0303	0.6013	0.0521	0.5233
(10, 500)	0.0690	0.0190	0.9517	0.0555	0.9417
(50, 500)	0.0594	0.0094	1	0.2252	1
(100, 500)	0.0537	0.0037	1	0.9757	1
(300, 500)	0.0515	0.0015	1	1	1

The powers are evaluated under the alternative hypothesis where $\mathbf{\Sigma} = \mathrm{diag}(1, 0.05, 0.05, 0.05, \ldots)$. From the simulation result, the behaviour of corrected likelihood ratio test becomes better with increasing dimension. On the other hand, the type I error of the traditional likelihood ratio test becomes worse with increasing dimension. It is also shown that once dimension is larger than 15, the corrected likelihood criterion has a satisfactory performance.

5.6.2 Correction of likelihood ratio test for equality: two-matrices case

Consider the likelihood ratio criterion for testing the equality of two population covariance matrices. In §5.3, we discuss the likelihood ratio criterion for testing the equality of q population covariance matrices and we have seen that its distribution can be decomposed into the product of $q-1$ mutually independent sub-criteria (see Theorem 5.1). Later, in Theorem 5.2, the likelihood ratio test can be further decomposed into the product of function of $pq-1$ Beta random variables. Although the precise distribution is known in theory as functions of Beta variables, its numerical evaluation is much too complex for practical applications especially for large p and q. Therefore, in this section, we provide the asymptotic distribution under the large dimensional scheme.

First, we introduce the correction of likelihood ratio criterion for testing the equality of two population covariance matrices. For the criterion V_1 defined in Eq. (5.8) with $q=2$, we have with $N_j = n_j - 1$, $j = 1, 2$ and $N = N_1 + N_2$,

$$V_1 = \frac{|\mathbf{A}_1|^{N_1/2}|\mathbf{A}_2|^{N_2/2}}{|\mathbf{A}_1 + \mathbf{A}_2|^{N/2}} = c_1^{\frac{1}{2}pN_1} c_2^{\frac{1}{2}pN_2} \cdot \frac{|\mathbf{S}_1\mathbf{S}_2^{-1}|^{N_1/2}}{|c_1\mathbf{S}_1\mathbf{S}_2^{-1} + c_2|^{N/2}}, \tag{5.37}$$

where $\mathbf{S}_j = \mathbf{A}_j/N_j$, $j = 1, 2$ are the unbiased sample covariance matrices and we have set $c_j = N_j/N$. We first consider the statistic

$$V_1^* = \frac{|\mathbf{S}_1\mathbf{S}_2^{-1}|^{N_1/2}}{|c_1\mathbf{S}_1\mathbf{S}_2^{-1} + c_2|^{N/2}}. \tag{5.38}$$

Theorem 5.5 *For the criterion V_1^* defined in Eq.* (5.38)*, assume the following large-dimensional scheme:*

$$y_{N_1} = \frac{p}{N_1} \to y_1 \in (0,1)$$
$$y_{N_2} = \frac{p}{N_2} \to y_2 \in (0,1)$$

Then we have

$$-\frac{2}{N}\log V_1^* - pd_2(y_{N_1}, y_{N_2}) \xrightarrow{\mathscr{D}} \mathcal{N}(\mu_2, \sigma_2^2), \tag{5.39}$$

where

$$\begin{aligned}
&N_j = n_j - 1\,, \quad j = 1, 2,\\
d_2(y_1, y_2) &= \frac{y_1 + y_2 - y_1y_2}{y_1y_2}\log\left(\frac{y_1 + y_2}{y_1 + y_2 - y_1y_2}\right)\\
&\quad + \frac{y_1(1-y_2)}{y_2(y_1+y_2)}\log(1-y_2) + \frac{y_2(1-y_1)}{y_1(y_1+y_2)}\log(1-y_1),\\
\mu_2 = \mu_2(y_1, y_2) &= \frac{1}{2}\log\left(\frac{y_1 + y_2 - y_1y_2}{y_1 + y_2}\right) - \frac{y_1\log(1-y_2) + y_2\log(1-y_1)}{y_1 + y_2},\\
\sigma_2^2 = \sigma_2^2(y_1, y_2) &= -\frac{2y_1^2\log(1-y_2) + y_2^2\log(1-y_1)}{(y_1+y_2)^2} - 2\log\left(\frac{y_1 + y_2}{y_1 + y_2 - y_1y_2}\right).
\end{aligned}$$

Moreover,

$$-\frac{2}{N}\log V_1 - p\tilde{d}_2(y_{N_1}, y_{N_2}) \xrightarrow{\mathscr{D}} \mathcal{N}(\mu_2, \sigma_2^2), \tag{5.40}$$

with

$$\tilde{d}_2(y_{N_1}, y_{N_2}) = d_2(y_{N_1}, y_{N_2}) - \frac{y_{N_1}}{y_{N_1}+y_{N_2}} \log \frac{y_{N_1}}{y_{N_1}+y_{N_2}} - \frac{y_{N_2}}{y_{N_1}+y_{N_2}} \log \frac{y_{N_2}}{y_{N_1}+y_{N_2}} \, .$$

Proof Let $\mathbf{F}_n = \mathbf{S}_1\mathbf{S}_2^{-1}$ be the Fisher random matrix associated to the sample covariance matrices $\mathbf{S}_j$'s. By Eq. (5.38),

$$-\frac{2}{N}\log V_1^* = \log|c_1\mathbf{F}_n + c_2| - \frac{N_1}{N}\log|\mathbf{F}_n| = p\int f(x)dF_n(x) \, ,$$

where $f(x) = \log(c_1x+c_2) - c_1\log(x)$ and F_n is the ESD of $\mathbf{F}_n$. Notice that $c_j = y_{N_j}/(y_{N_1} + y_{N_2})$ so that f can be written as

$$f(x) = \log(y_{N_1} + y_{N_2}x) - \frac{y_{N_2}}{y_{N_1}+y_{N_2}}\log x - \log(y_{N_1}+y_{N_2}) \, .$$

By the two-sample substitution principle in Theorem 3.17, the CLT for $-\frac{2}{N}\log V_1^*$ is given in Theorem 3.10 where the centring parameter is to be calculated as

$$\int f(x)dF_{y_{N_1},y_{N_2}}(x) \, ,$$

while the limiting parameters μ_2 and σ_2^2 are evaluated with respect to the LSD F_{y_1,y_2}. The value of the centring parameter can be calculated as in Lemma 2.26 and it equals $d_2(y_{N_1}, y_{N_2})$. The values of μ_2 and σ_2^2 are derived using the results given in Example 3.12.

The last conclusion follows from Eq. (5.37) and the simple fact that $c_j = y_{N_j}/(y_{N_1}+y_{N_2})$. The proof is complete. □

Again we see by its proof that Theorem 5.5 is also valid for non normal populations that respect the assumptions used in Theorem 3.10. In this case, the centring parameter $d_2(y_{N_1}, y_{N_2})$ remains the same while the limiting parameters μ_2 and σ_2^2 are to be adapted according to general results given in Example 3.12.

Using the theorem, a test can be designed with critical region

$$-\frac{2}{N}\log V_1 \geq p\tilde{d}_2(y_{N_1}, y_{N_2}) + \mu_2 + \sigma_2 z_\alpha \, ,$$

where z_α is the αth upper quantile of standard normal. This test is called the *corrected likelihood ratio test* for the equality hypothesis between two covariance matrices.

For different values of (p, n_1, n_2), empirical sizes and powers of the traditional LR criterion based on V_1 with asymptotic distribution given in (5.18) and the corrected LR criterion are evaluated using simulation with 10,000 independent replications. The nominal test level is 0.05 and real Gaussian variables are used. Results are summarised in Table 5.2. As we can see, when the dimension p increases, the traditional LR criterion leads to a dramatically high test size while the corrected LR criterion remains accurate. Furthermore, for moderate dimensions like $p = 20$ or 40, the sizes of the traditional LR criterion are much higher than 5%, whereas the ones of the corrected LR criterion are very close. By a closer look at the column showing the difference with 5%, we note that this difference rapidly decreases as p increases for the corrected criterion. Next, empirical powers are evaluated under the alternative hypothesis $\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-1} = \text{diag}(3, 1, 1, 1, \ldots)$. From simulation result, it is suggested that once the dimension is larger than 10, the corrected large dimensional criterion should be applied.

Table 5.2 *Comparison of Type I error and power between corrected likelihood ratio test and traditional likelihood ratio test*

	Corrected likelihood ratio test			Traditional likelihood ratio test	
(p, n_1, n_2)	Type I error	5% difference	Power	Type I error	Power
	$(y_1, y_2) = (0.05, 0.05)$				
(5, 100, 100)	0.0770	0.0270	1	0.0582	1
(10, 200, 200)	0.0680	0.0180	1	0.0684	1
(20, 400, 400)	0.0593	0.0093	1	0.0872	1
(40, 800, 800)	0.0526	0.0026	1	0.1339	1
(80, 1600, 1600)	0.0501	0.0001	1	0.2687	1
(160, 3200, 3200)	0.0491	-0.0009	1	0.6488	1
(320, 6400, 6400)	0.0447	-0.0053	0.9671	1	1
	$(y_1, y_2) = (0.05, 0.1)$				
(5, 100, 50)	0.0781	0.0281	0.9925	0.0640	0.9849
(10, 200, 100)	0.0617	0.0117	0.9847	0.0752	0.9904
(20, 400, 200)	0.0573	0.0073	0.9775	0.1104	0.9938
(40, 800, 400)	0.0561	0.0061	0.9765	0.2115	0.9975
(80, 1600, 800)	0.0521	0.0021	0.9702	0.4954	0.9998
(160, 3200, 1600)	0.0520	0.0020	0.9702	0.9433	1
(320, 6400, 3200)	0.0510	0.0010	1	0.9939	1

5.6.3 Correction of likelihood ratio test for equality: multiple-matrices case

In section 5.2, the criterion of likelihood ratio for testing the equality of q population covariance matrix is derived in Eq. (5.6). In Eq. (5.8), the corresponding Bartlett correction is given. Similar to the discussion in the previous subsection, the exact distribution of the likelihood criterion is known to be the one of a product of independent Beta random variables and this distribution is far too complex for practical use, especially when p or q is large. We provide below an asymptotic distribution under the large-dimensional setting.

Theorem 5.6 *Consider the criterion $V_1 = V_{12} \times cdots \times V_{1q}$ defined in Eq. (5.8) and Eq.* (5.16) *and assume the following large dimensional scheme: for any $g = 2, \cdots, q$*

$$y_{n1}^{(g)} = \frac{p}{N_1 + \cdots + N_{g-1}} \to y_1^{(g)} \in (0,1),$$

$$y_{n2}^{(g)} = \frac{p}{N_g} \to y_2^{(g)} \in (0,1).$$

Therefore,

$$\sum_{g=2}^{q} \left\{ -\frac{2}{N_1 + \cdots + N_g} \log V_{1g} - p\tilde{d}_2(y_{n1}^{(g)}, y_{n2}^{(g)}) \right\} \xrightarrow{\mathscr{D}} \mathcal{N}\left(\sum_{g=2}^{q} \mu_2(y_1^{(g)}, y_2^{(g)}), \sum_{g=2}^{q} \sigma_2^2(y_1^{(g)}, y_2^{(g)}) \right), \tag{5.41}$$

where $\tilde{d}_2, \mu_2, \sigma_2^2$ are the functions defined in Theorem 5.5.

Proof According to Theorem 5.1, the factors V_{1g}'s in the decomposition of V_1 are mutually independent, and each of them is distributed similarly to V_1 for two populations studied in Theorem 5.5. Therefore, the conclusion follows. □

We notice that it is unclear whether the conclusion of Theorem 5.6 is still valid when the data has a non-normal distribution. Of course, we still have $V_1 = \prod_{g=2}^{q} V_{1g}$, and under the condition of 4-th moments,

$$-\frac{2}{N_1+\cdots+N_g}\log V_{1g}-p\tilde{d}_2(y_{n1}^{(g)},y_{n2}^{(g)})\xrightarrow{\mathscr{D}}\mathcal{N}(\mu_2(y_1^{(g)},y_2^{(g)}),\sigma_2^2(y_1^{(g)},y_2^{(g)})).$$

But we do not know whether the V_{1g}'s are asymptotically independent for non-normal populations. It is conjectured that this asymptotic independence do hold but a rigorous proof is still needed.

5.6.4 Correction of likelihood ratio test for equality of several normal distributions

A problem close to the previous section is to test the equality of q normal distributions. The Bartlett corrected likelihood ratio test of the hypothesis is given in equation (5.15). Theorem 5.3 proves that V_1 and V_2 are independent. Note that $\log V = \log V_1 + \log V_2$. To find the limit distribution of $\log V$, we only need to find the limit distribution of $\log V_2$. With the definition of V_2 (see equation (5.15)), we have

$$\log V_2 = -\log|\mathbf{I}+\frac{n-q}{q-1}\mathbf{F}|,$$

where $\mathbf{F} = (q-1)^{-1}(\mathbf{B}-\mathbf{A})(N^{-1}\mathbf{A})^{-1}$ is multivariate F matrix with degree of freedom $[(q-1),N]$. Hence we have the following theorem.

Theorem 5.7 *If*

$$\begin{aligned}y_{n1}^{(1)} &= \frac{p}{q-1}\to y_1^{(1)}>0\\ y_{n2}^{(1)} &= \frac{p}{n-q}\to y_2^{(1)}\in(0,1),\end{aligned}$$

Then

$$\log V_2 - pd_3(y_{n1}^{(1)},y_{n2}^{(1)})\xrightarrow{\mathscr{D}}\mathcal{N}\left[\mu_3(y_1^{(1)},y_2^{(1)}),\sigma_3^2(y_1^{(1)},y_2^{(1)})\right], \tag{5.42}$$

where

$$\begin{aligned}
d_3(y_1,y_2) &= \frac{1-y_2}{y_2}\log\left(\frac{\alpha}{1-y_2}\right)-\frac{y_1+y_2}{y_1y_2}\log\left(\frac{h\alpha-y_2\beta}{h(1-y_2)}\right)\\
&\quad+\begin{cases}\frac{1-y_1}{y_1}\log\left(\frac{\alpha-h\beta}{1-y_2}\right) & \textit{if } y_1\in(0,1),\\ \frac{y_1-1}{y_1}\log\left(\frac{\alpha-h^{-1}\beta}{1-y_2}\right) & \textit{if } y_1\geq 1,\end{cases}\\
\mu_3 &= \frac{1}{2}\log\left(\frac{(\alpha^2-\beta^2)h^2}{(h\alpha-y_2\beta)^2}\right),\\
\sigma_3^2 &= 2\log\left(\frac{\alpha^2}{\alpha^2-\beta^2}\right),\\
c &= y_1/y_2,\\
\alpha &= \frac{1}{2}\left[\sqrt{(1-y_2)^2+c(1+h)^2}+\sqrt{(1-y_2)^2+c(1-h)^2}\right],\\
\beta &= \frac{1}{2}\left[\sqrt{(1-y_2)^2+c(1+h)^2}-\sqrt{(1-y_2)^2+c(1-h)^2}\right].
\end{aligned}$$

Proof First we apply Theorem 2.23 to calculate d_3. We use $f(x) = -\log(1 + cx) = \log\left(\frac{|\alpha+z\beta|^2}{(1-y_2)^2}\right)$. Then similar to the calculations in Lemma 2.26, we have

$$\begin{aligned} d_3(y_1, y_2) &= \frac{h^2(1-y_2)}{4\pi i} \oint_{|z|=1} \frac{\log\left(1 + c\frac{|1+hz|^2}{(1-y_2)^2}\right)(1-z^2)^2 dz}{z(1+hz)(z+h)(y_2+hz)(y_2z+h)} \\ &= \frac{h^2(1-y_2)}{4\pi i} \oint_{|z|=1} \frac{\log\left(\frac{|\alpha+\beta z|^2}{(1-y_2)^2}\right)(1-z^2)^2 dz}{z(1+hz)(z+h)(y_2+hz)(y_2z+h)} \\ &= \frac{h^2(1-y_2)}{2\pi i} \oint_{|z|=1} \frac{\log\left(\frac{\alpha+\beta z}{1-y_2}\right)(1-z^2)^2 dz}{z(1+hz)(z+h)(y_2+hz)(y_2z+h)} \\ &= \frac{1-y_2}{y_2} \log\left(\frac{\alpha}{1-y_2}\right) - \frac{y_1+y_2}{y_1y_2} \log\left(\frac{h\alpha - y_2\beta}{h(1-y_2)}\right) \\ &\quad + \begin{cases} \frac{1-y_1}{y_1} \log\left(\frac{\alpha-h\beta}{1-y_2}\right) & \text{if } y_1 \in (0,1), \\ \frac{y_1-1}{y_1} \log\left(\frac{\alpha-h^{-1}\beta}{1-y_2}\right) & \text{if } y_1 \geq 1. \end{cases} \end{aligned}$$

Next, we use Theorem 3.10 to calculate μ_3 and σ_3^2. We have

$$\begin{aligned} \mu_3 &= -\lim_{r\uparrow 1} \frac{1}{2\pi i} \oint_{|z|=1} \log(|\alpha + z\beta|^2) \left[\frac{z}{z^2 - r^2} - \frac{1}{z + y_2/h}\right] dz \\ &= \frac{1}{2} \log\left(\frac{(\alpha^2 - \beta^2)h^2}{(h\alpha - y_2\beta)^2}\right), \\ \sigma_3^2 &= -\lim_{r\uparrow 1} \frac{1}{2\pi^2} \oint \oint \frac{\log(|\alpha + z_1\beta|^2 \log(|\alpha + z_2\beta|^2}{(z_1 - rz_2)^2} dz_1 dz_2 \\ &= 2\log\left(\frac{\alpha^2}{\alpha^2 - \beta^2}\right). \end{aligned}$$

□

Corollary 5.8 *Under the conditions of Theorems 5.6 and 5.7, we have*

$$\log W - p \sum_{g=1}^{q} d_2(y_{n1}^{(g)}, y_{n2}^{(g)}) \xrightarrow{\mathscr{D}} N\left(\sum_{g=1}^{q} \mu_2(y_1^{(g)}, y_2^{(g)}), \sum_{g=1}^{q} \sigma_2^2(y_1^{(g)}, y_2^{(g)})\right). \tag{5.43}$$

5.6.5 A large-dimension trace criterion for testing equality of several normal distributions

With reference to various results found in the previous section, we notice that $y_1^{(g)}, y_2^{(g)}$, $g \geq 2$ cannot be too close to 1. Otherwise, $\mu_2^{(g)}, \sigma_2^2$ will become unstable. Likely, $y_2^{(1)}$ cannot be too close to 1, otherwise the limiting parameters μ_3 and σ_3^2 become unstable. However, μ_3 and σ_3^2 are still well defined when $y_1^{(1)}$ equals to or is larger than 1. In such situations, to reduce the drawback, Nagao's trace criterion introduced in §5.2 (see Eq. (5.10)) is a more suitable solution to the test problem. In the classical setting with p fixed, it is proved that the asymptotic distribution of the criterion is χ_f^2 with degree of freedom $f = \frac{1}{2}(q-1)p(p+1)$.

To introduce large-dimensional correction, we first consider the case of $q = 2$. Recall the notations $n = n_1 + n_2$, $N_j = n_j - 1$, $j = 1, 2$ and $N = N_1 + N_2$. The Nagao's criterion is now

$$\begin{aligned}
&N_1 \mathrm{tr}(\mathbf{S}_1\mathbf{S}^{-1} - \mathbf{I})^2 + N_2 \mathrm{tr}(\mathbf{S}_2\mathbf{S}^{-1} - \mathbf{I})^2 \\
&= n\lambda \mathrm{tr}(\mathbf{F}(\lambda\mathbf{F} + (1-\lambda)\mathbf{I})^{-1} - \mathbf{I})^2 + n(1-\lambda)\mathrm{tr}((\lambda\mathbf{F} + (1-\lambda)\mathbf{I})^{-1} - \mathbf{I})^2 \\
&= 2n\lambda(1-\lambda)\mathrm{tr}(\mathbf{F} - \mathbf{I})^2(\lambda\mathbf{F} + (1-\lambda)\mathbf{I})^{-2},
\end{aligned}$$

where $\lambda = \frac{N_1}{N}$ and $\mathbf{F} = \mathbf{S}_1\mathbf{S}_2^{-1}$. Therefore, the test using $\mathrm{tr}(\mathbf{F} - \mathbf{I})^2(\lambda\mathbf{F} + (1-\lambda)\mathbf{I})^{-2}$ is equivalent to Nagao's test. We set the following

Theorem 5.9 *Assume*

$$\begin{aligned}
y_{N_1} &= \frac{p}{N_1} \to y_1 > 0 \\
y_{N_2} &= \frac{p}{N_2} \to y_2 \in (0, 1).
\end{aligned}$$

Then

$$tr(\mathbf{F} - \mathbf{I})^2(\lambda\mathbf{F} + (1-\lambda)\mathbf{I})^{-2} - pd_5(y_{n1}, y_{n2}) \xrightarrow{\mathscr{D}} \mathcal{N}(\mu_5, \sigma_5^2), \tag{5.44}$$

where

$$\begin{aligned}
d_5(y_1, y_2) &= y_1 + y_2, \\
\mu_5 &= y_1 + y_2 + 2y_1y_2, \\
\sigma_5^2 &= 8((y_1 + y_2)^2 + 2(y_1 + y_2)(y_1^2 + y_2^2 - y_1y_2) + y_1y_2(2y_1 - y_2)(2y_2 - y_1)).
\end{aligned}$$

Proof The proof is based on the CLT in Theorem 3.10. Note that $\lambda = \frac{y_{N_2}}{y_{N_1} + y_{N_2}}$. Let $f(x) = \frac{(x-1)^2}{(\lambda x + 1 - \lambda)^2} = \frac{(x-1)^2(y_{N_1} + y_{N_2})^2}{(y_{N_2}x + y_{N_1})^2}$.

For the calculation of d_5 and to simplify the notation, we denote (y_{N_1}, y_{N_2}) simply by (y_1, y_2). Applying Theorem 2.23, we need to convert x as

$$x = \frac{|1 + hz|^2}{(1 - y_2)^2} = \frac{(1 + hz)(1 + hz^{-1})}{(1 - y_2)^2}, \quad |z| = 1.$$

Then

$$\begin{aligned}
f(x) &= (y_1 + y_2)^2 \frac{[(1 + hz)(z + h) - z(1 - y_2)^2]^2}{[y_2(1 + hz)(h + z) + zy_1(1 - y_2)^2]^2} \\
&= (y_1 + y_2)^2 \frac{[(1 + hz)(z + h) - z(1 - y_2)^2]^2}{(y_2 + hz)^2(h + zy_2)^2}.
\end{aligned}$$

We use the following equation to calculate d_5:

$$\begin{aligned}
&d_5(y_1, y_2) \\
&= \frac{h^2(1 - y_2)}{-4\pi i} \oint_{|z|=1} \frac{[(1 + hz)(z + h) - z(1 - y_2)^2]^2(y_1 + y_2)^2(1 - z^2)^2}{z(1 + hz)(h + z)(y_2z + h)^3(y_2 + hz)^3} dz.
\end{aligned}$$

The function under the integral has three poles in the unit disk: two simple poles 0 and $-h$, and one third-order pole $-y_2/h$. The residues of the two simple poles are respectively,

$$\frac{(1 - y_2)(y_1 + y_2)^2}{y_2^3}, \quad \text{and} \quad \frac{(1 - y_1)(y_1 + y_2)^2}{y_1^3}.$$

The residue of $-y_2/h$ is half of the second-order derivative of the integrand function multiplied by $(z + y_2/h)^3$, which is

$$
\begin{aligned}
&\frac{h^2(y_1+y_2)^2(1-y_2)[(1-y_2)(h-y_2/h)+(y_2/h)(1-y_2)^2]^2(1-y_2^2/h^2)^2}{2(-y_2/h)(1-y_2)(h-y_2/h)h^3(h-y_2^2/h)^3}\times\\
&\bigg\{\bigg[\frac{-4y_2+2(1+h^2)-2(1-y_2)^2}{(1-y_2)(h-y_2/h)+y_2(1-y_2)^2/h}+\frac{4y_2/h}{1-y_2^2/h^2}+\frac{1}{y_2/h}-\frac{h}{1-y_2}\\
&-\frac{1}{h-y_2/h}-\frac{3y_2}{h-y_2^2/h}\bigg]^2+\bigg[\frac{4h}{(1-y_2)(h-y_2/h)+y_2(1-y_2)^2/h}\\
&-\frac{2[-2y_2+(1+h^2)-(1-y_2)^2]^2}{[(1-y_2)(h-y_2/h)+y_2(1-y_2)^2/h]^2}-\frac{4}{1-y_2^2/h^2}-\frac{8y_2^2/h^2}{(1-y_2^2/h^2)^2}\\
&+\frac{1}{(y_2/h)^2}+\frac{h^2}{(1-y_2)^2}+\frac{1}{(h-y_2/h)^2}+\frac{3y_2^2}{(h-y_2^2/h)^2}\bigg]\bigg\}\\
=&-\frac{(y_1+y_2)[(y_1+y_2)^2(y_1^2+y_2^2-y_1y_2)-y_1y_2(y_1+y_2)(y_1^2+y_2^2)+2y_1^3y_2^3]}{y_1^3y_2^3}.
\end{aligned}
$$

Combining the above results leads to $d_5(y_1, y_2) = -\frac{1}{2}$(sum of residues)$= y_1 + y_2$.

The limiting parameters μ_5 and σ_5^2 are calculated using Theorem 3.10 as follows. First for the asymptotic mean,

$$
\begin{aligned}
\mu_5 =& \lim_{r\uparrow 1}\frac{(y_1+y_2)^2}{2\pi i}\oint_{|z|=1}\frac{[(1+hz)(z+h)-z(1-y_2)^2]^2}{(y_2+hz)^2(h+zy_2)^2}\\
&\left[\frac{z}{z^2-r^2}-\frac{1}{z+y_2/h}\right]dz\\
=&\frac{1}{2}(y_1+y_2)^2\bigg[\frac{[(1+h)^2-(1-y_2)^2]^2}{(h+y_2)^4}+\frac{[(1-h)^2-(1-y_2)^2]^2}{(h-y_2)^4}\\
&+\frac{[(1-y_2)(h-y_2/h)+y_2/h(1-y_2)^2]^2(-2y_2/h)}{h^2(h-y_2^2/h)^2(y_2^2/h^2-1)}\times\\
&\bigg(\frac{-4y_2+2(1+h^2)-2(1-y_2)^2}{(1-y_2)(h-y_2/h)+y_2/h(1-y_2)^2}-\frac{1}{y_2/h}+\frac{2y_2/h}{y_2^2/h^2-1}-\frac{2y_2}{h-y_2^2/h}\bigg)\\
&-\frac{2[(1-y_2)(h-y_2/h)+y_2/h(1-y_2)^2]^2}{2h^2(h-y_2^2/h)^2}\times\bigg(\\
&\bigg(\frac{-4y_2+2(1+h^2)-2(1-y_2)^2}{(1-y_2)(h-y_2/h)+y_2/h(1-y_2)^2}-\frac{2y_2}{h-y_2^2/h}\bigg)^2\\
&+\frac{4h}{(1-y_2)(h-y_2/h)+y_2/h(1-y_2)^2}-\frac{2[-2y_2+(1+h^2)-(1-y_2)^2]^2}{((1-y_2)(h-y_2/h)+y_2/h(1-y_2)^2)^2}\\
&+\frac{2y_2^2}{(h-y_2^2/h)^2}\bigg)\bigg]\\
=& y_1+y_2+2y_1y_2.
\end{aligned}
$$

And for the asymptotic variance,

$$
\sigma_5^2 = -\lim_{r\uparrow 1}\frac{(y_1+y_2)^4}{\pi^2}\oint\oint_{|z_1|=|z_2|=1}\frac{[(1+hz_1)(z_1+h)-z_1(1-y_2)^2]^2}{(y_2+hz_1)^2(h+z_1y_2)^2}\times
$$

$$\frac{[(1+hz_2)(z_2+h)-z_2(1-y_2)^2]^2}{(y_2+hz_2)^2(h+z_2y_2)^2}\frac{1}{(z_1-rz_2)^2}dz_1dz_2$$

$$=\frac{2(y_1+y_2)^4(1-y_2)^2h}{\pi i}\oint_{|z_1|=1}\frac{[(1+hz_1)(z_1+h)-z_1(1-y_2)^2]^2}{(y_2+hz_1)^4(h+z_1y_2)^2}\times$$

$$\Big(\frac{2y_1}{(y_1+y_2)(1-y_2)}+\frac{2}{hz_1+y_2}\Big)dz_1$$

$$=8((y_1+y_2)^2+2(y_1+y_2)(y_1^2+y_2^2-y_1y_2)+y_1y_2(2y_1-y_2)(2y_2-y_1)).$$

The proof is complete. □

Consider next the general case with more than two populations. It is unclear how to extend Nagao's criterion Eq.(5.10) to this general situation with large dimensional data. However, we can introduce a similar criterion as follow:

$$T_q=\sum_{k=2}^{q}\lambda_k\operatorname{tr}(\mathbf{F}_k-1)^2(y_{nk2}\mathbf{F}_k+y_{nk1}\mathbf{I})^{-2},$$

where λ_k are some positive weights, and

$$\mathbf{F}_k=\mathbf{S}_k\underline{\mathbf{S}}_{k-1}^{-1},$$

$$\underline{\mathbf{S}}_k=\frac{1}{N_1+\cdots+N_k}(\mathbf{A}_1+\cdots+\mathbf{A}_k),$$

$$y_{nk2}=\frac{p}{N_k},\quad y_{nk1}=\frac{p}{N_1+\cdots+N_{k-1}}.$$

For normal populations, $\mathbf{F}_2,\cdots,\mathbf{F}_q$ are mutually independent multivariate F-matrices. Since each term in T_q has the form of the statistic studied in Theorem 5.9, a CLT for T_q is readily found as follows.

Corollary 5.10 *Assume that $y_{nkj}\to y_{kj}>0$, $k=2,\cdots,q$, $j=1,2$ with $y_{k1}\in(0,1)$. Then*

$$T_q-p\sum_{k=2}^{q}d_5\lambda_k(y_{nk1},y_{nk2})\xrightarrow{\mathscr{D}}\mathcal{N}\left(\sum_{k=2}^{q}\lambda_k\mu_5(y_{k1},y_{k2}),\sum_{k=2}^{q}\lambda_k^2\sigma_5^2(y_{k1},y_{k2})\right),\qquad(5.45)$$

where the functions μ_5 and σ_5^2 are given in Theorem 5.9.

5.7 Large-dimensional sphericity test

Consider a sample $\mathbf{x}_1,\ldots,\mathbf{x}_n$ from a p-dimensional multivariate distribution with covariance matrix Σ_p and mean $\boldsymbol{\mu}$. The sample covariance matrix is

$$\mathbf{S}_n=\frac{1}{N}\sum_{j=1}^{n}(\mathbf{x}_j-\bar{\mathbf{x}})(\mathbf{x}_j-\bar{\mathbf{x}})^*,\quad\bar{\mathbf{x}}=\frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_i\,.\qquad(5.46)$$

Let (ℓ_j) denote its eigenvalues.

The likelihood ratio criterion for the test is given in (5.26) of §5.4, namely

$$\lambda=\left(\frac{(\ell_1\cdots\ell_p)^{1/p}}{\frac{1}{p}(\ell_1+\cdots+\ell_p)}\right)^{\frac{1}{2}pn},$$

which is a power of the ratio of the geometric mean of the sample eigenvalues to the arithmetic mean. It is here noticed that in this formula it is necessary to assume that $p \leq n$ to avoid null eigenvalues in (the numerator of) λ. If we let $n \to \infty$ while keeping p fixed, classical asymptotic theory shows that under the null hypothesis, $-2 \log \lambda \xrightarrow{\mathscr{D}} \chi^2_f$, a chi-square distribution with degree of freedom $f = \frac{1}{2}p(p+1) - 1$. This asymptotic distribution is further refined by the following Box-Bartlett correction (referred as *BBLRT*), see Eq. (5.27),

$$\mathbb{P}(-2\rho \log \lambda \leq x) = \varphi_f(x) + \omega_2 \left\{\varphi_{f+4}(x) - \varphi_f(x)\right\} + O(n^{-3}) \,, \tag{5.47}$$

where $\varphi_m(x) = P(\chi^2_m \leq x)$ and

$$\rho = 1 - \frac{2p^2 + p + 2}{6pN}, \qquad \omega_2 = \frac{(p+2)(p-1)(p-2)(2p^3 + 6p^2 + 3p + 2)}{288p^2N^2\rho^2} \,.$$

By observing that the asymptotic variance of $-2 \log \lambda$ is proportional to $\mathrm{tr}\{\mathbf{\Sigma}(\mathrm{tr}\,\mathbf{\Sigma})^{-1} - p^{-1}\mathbf{I}_p\}^2$, John (1971) proposed to use the following statistic, see Eq. (5.28),

$$T_2 = \frac{p^2 n}{2} \,\mathrm{tr}\left\{\mathbf{S}_n(\mathrm{tr}\,\mathbf{S}_n)^{-1} - p^{-1}\mathbf{I}_p\right\}^2$$

for testing sphericity. When p is fixed and $n \to \infty$, under the null hypothesis, it also holds that $T_2 \xrightarrow{\mathscr{D}} \chi^2_f$. The criterion based on this χ^2 limiting is referred to as *John's test.* It is observed that T_2 is proportional to the square of the coefficient of variation of the sample eigenvalues, namely

$$T_2 = \frac{np}{2} \cdot \frac{p^{-1} \sum(\ell_i - \overline{\ell})^2}{\overline{\ell}^2} \,, \qquad \text{with } \overline{\ell} = \frac{1}{n}\sum_i \ell_i \,.$$

Following the idea of the Box-Bartlett correction, Nagao (1973a) established an expansion for the distribution function of the statistics T_2,

$$\begin{aligned}\mathbb{P}(T_2 \leq x) = \varphi_f(x) + \frac{1}{n}&\left\{a_p\varphi_{f+6}(x) + b_p\varphi_{f+4}(x) + c_p\varphi_{f+2}(x) + d_p\varphi_f(x)\right\} \\ &+ O(n^{-2}),\end{aligned} \tag{5.48}$$

where

$$\begin{aligned} a_p &= \frac{1}{12}(p^3 + 3p^2 - 12 - 200p^{-1}), \; b_p = \frac{1}{8}(-2p^3 - 5p^2 + 7p - 12 - 420p^{-1}) \,, \\ c_p &= \frac{1}{4}(p^3 + 2p^2 - p - 2 - 216p^{-1}) \,, d_p = \frac{1}{24}(-2p^3 - 3p^2 + p + 436p^{-1}) \,. \end{aligned}$$

The criterion based on this expansion is referred to as *Nagao's test.*

As discussed in previous chapters, classical multivariate procedures are in general biased with large-dimensional data. It is again confirmed here by a small simulation experiment that explores the performance of the BBLRT and Nagao's test with growing dimension p. The sample size is set to $n = 64$ while dimension p increases from 4 to 60 (other experiments with larger sample sizes n lead to very similar conclusions), and the nominal significance level is $\alpha = 0.05$. The samples come from normal vectors with mean zero and identity covariance matrix, and each pair of (p, n) is assessed with 10000 independent replications.

Table 5.3 gives the empirical sizes of BBLRT and Nagao's test. It is found here that when the dimension to sample size ratio p/n is below $1/2$, both tests have an empirical size close to the nominal test level 0.05. Then when the ratio grows up, the BBLRT becomes quickly biased while Nagao's test still has a correct empirical size. It is striking that although Nagao's test is derived under classical "p fixed, $n \to \infty$" regime, it is remarkably robust against dimension inflation.

Table 5.3 *Empirical sizes of BBLRT and Nagao's test at* 5% *significance level based on* 10000 *independent replications using normal vectors* $\mathcal{N}(0, \mathbf{I}_p)$ *for* $n = 64$ *and different values of p.*

(p, n)	(4,64)	(8,64)	(16,64)	(32,64)	(48,64)	(56,64)	(60,64)
BBLRT	0.0483	0.0523	0.0491	0.0554	0.1262	0.3989	0.7605
Nagao's test	0.0485	0.0495	0.0478	0.0518	0.0518	0.0513	0.0495

In this section, novel corrections to both LRT and John's test are proposed to cope with large-dimensional data. These corrections are applicable to non-normal populations and it will be shown that John's test is robust against the inflation of the dimension, i.e. its limiting distribution under the large-dimensional scheme coincides with its limiting distribution derived under the classical low-dimensional scheme.

It is thus not assumed anymore that the population is normal. The data structure is as follows. The observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ have the representation $\mathbf{x}_j = \mathbf{\Sigma}_p^{1/2} X_j$ where the $p \times n$ table $\{X_1, \ldots, X_n\} = \{x_{ij}\}_{1 \le i \le p, 1 \le j \le n}$ are made with an array of i.i.d. standardised random variables (mean 0 and variance 1). This setting, quite general, has been e.g. already used in Chapter **??** and is motivated by the random matrix theory. Furthermore, under the null hypothesis $H_0 : \mathbf{\Sigma}_p = \sigma^2 \mathbf{I}_p$ (σ^2 is unspecified), we notice that both LRT and John's test are independent of the scale parameter σ^2 under the null. Therefore, we can assume w.l.o.g. $\sigma^2 = 1$ when dealing with the null distributions of these test statistics. This will be assumed in all the sections.

Similar to the CLT's in Chapter 3, an indicator κ is set to 2 when $\{x_{ij}\}$ are real-valued and to 1 when they are complex-valued. Let $\beta = E|x_{ij}|^4 - 1 - \kappa$ be the fourth cumulant of the variables for both cases. Note that for normal variables, $\beta = 0$ (recall that for a standard complex-valued normal random variable, its real and imaginary parts are two iid. $\mathcal{N}(0, \frac{1}{2})$ real random variables).

5.7.1 The corrected likelihood ratio test (CLRT)

For the correction of LRT, let $\mathcal{L}_n = -2n^{-1} \log \lambda$ be the test statistic for $n \ge 1$.

Theorem 5.11 *Assume $\{x_{ij}\}$ are iid, satisfying $Ex_{ij} = 0, E|x_{ij}|^2 = 1, E|x_{ij}|^4 < \infty$. Then under H_0 and when $\frac{p}{N} = y_N \to y \in (0, 1)$,*

$$\mathcal{L}_n + (p - N) \cdot \log(1 - \frac{p}{N}) - p \xrightarrow{\mathscr{D}} \mathcal{N}\left\{-\frac{\kappa - 1}{2} \log(1 - y) + \frac{1}{2}\beta y, -\kappa \log(1 - y) - \kappa y\right\}. \tag{5.49}$$

Proof Recall that for the Marčenko-Pastur distribution F_y of index $y < 1$, the following integrals are calculated in Example 2.12:

$$F^y(\log x) = \frac{y-1}{y}\log(1-y) - 1\,, \quad F^y(x) = 1.$$

Let $A_n \equiv \sum_{i=1}^p \log \ell_i - pF^{y_N}(\log x)$ and $B_n \equiv \sum_{i=1}^p \ell_i - pF^{y_N}(x)$. By the one-sample substitution principle in Theorem 3.16 and Proposition 3.8, we have that

$$\begin{pmatrix} A_n \\ B_n \end{pmatrix} \xrightarrow{\mathscr{D}} \mathcal{N}(\mu_1, V_1).$$

Here the limiting parameters (μ_1, V_1) are calculated in Proposition 3.8 and they equal

$$\mu_1 = \begin{pmatrix} \frac{\kappa-1}{2}\log(1-y) - \frac{1}{2}\beta y \\ 0 \end{pmatrix},$$

and

$$V_1 = \begin{pmatrix} -\kappa\log(1-y) + \beta y & (\beta+\kappa)y \\ (\beta+\kappa)y & (\beta+\kappa)y \end{pmatrix}.$$

Consequently, $-A_n + B_n$ is asymptotically normal with mean $-\frac{\kappa-1}{2}\log(1-y) + \frac{1}{2}\beta y$ and variance

$$V_1(1,1) + V_1(2,2) - 2V_1(1,2) = -\kappa\log(1-y) - \kappa y.$$

Besides,

$$\begin{aligned} \mathcal{L}_n &= -\Sigma_{i=1}^p \log \ell_i + p\log(\frac{1}{p}\Sigma_{i=1}^p \ell_i) \\ &= -(A_n + pF^{y_N}(\log x)) + p\log(\frac{1}{p}(B_n + p)) \\ &= -A_n - pF^{y_N}(\log x) + p\log(1 + \frac{B_n}{p}). \end{aligned}$$

Since $B_n \xrightarrow{\mathscr{D}} \mathcal{N}(0, y(\beta+\kappa))$, $B_n = O_p(1)$ and $\log(1+B_n/p) = B_n/p + O_p(1/p^2)$. Therefore,

$$\mathcal{L}_n = -A_n - pF^{y_N}(\log x) + B_n + O_p(\frac{1}{p})\,.$$

The conclusion then follows and the proof is complete. □

The test based on this asymptotic normal distribution is referred as the *corrected likelihood-ratio test* (*CLRT*). One may observe that the limiting distribution of the test crucially depends on the limiting dimension-to-sample-size ratio y through the factor $-\log(1-y)$. In particular, the asymptotic variance will blow up quickly when y approaches 1, so it is expected that the power will seriously break down. Monte-Carlo experiments in §5.7.3 will provide more details on this behaviour.

5.7.2 The corrected John's test (CJ)

Earlier than the asymptotic expansion (5.48) given in Nagao (1973a), John (1971) proved that when the observations are normal, the sphericity test based on T_2 is a locally most powerful invariant test. It is also established in John (1972) that under these conditions, the limiting distribution of T_2 under H_0 is χ^2_f with degree of freedom $f = \frac{1}{2}p(p+1) - 1$, or equivalently,

$$NU - p \xrightarrow{\mathscr{D}} \frac{2}{p}\chi^2_f - p \, ,$$

where for convenience, we have let $U = 2(np)^{-1}T_2$ and recall that $N = n - 1$ is the degree of freedom of the sample covariance matrix $\mathbf{S}_n$. This limit is established for $n \to \infty$ and a fixed dimension p. However, if we now let $p \to \infty$ in the right-hand side of the above result, it is not hard to see that $\frac{2}{p}\chi^2_f - p$ will tend to the normal distribution $\mathcal{N}(1, 4)$. It then seems "natural" to conjecture that when both p and n grow to infinity in some "proper" way, it may happen that

$$NU - p \xrightarrow{\mathscr{D}} \mathcal{N}(1, 4) \, . \tag{5.50}$$

Theorem 5.12 *Assume $\{x_{ij}\}$ are iid, satisfying $Ex_{ij} = 0, E|x_{ij}|^2 = 1, E|x_{ij}|^4 < \infty$, and let $U = 2(np)^{-1}T_2$ be the test statistic. Then under H_0 and when $p \to \infty, n \to \infty, \frac{p}{N} = y_N \to y \in (0, \infty)$,*

$$NU - p \xrightarrow{\mathscr{D}} \mathcal{N}(\kappa + \beta - 1, 2\kappa) \, . \tag{5.51}$$

The proof of Theorem 5.12 is based on the following lemma.

Lemma 5.13 *Let $\{\ell_i\}_{1 \le i \le p}$ be the eigenvalues of the sample covariance matrix $\mathbf{S}_n$. Then under H_0 and the conditions of Theorem 5.12, we have*

$$\begin{pmatrix} \sum_{i=1}^p \ell_i^2 - p(1 + y_n) \\ \sum_{i=1}^p \ell_i - p \end{pmatrix} \xrightarrow{\mathscr{D}} \mathcal{N}(\mu_2, V_2),$$

with

$$\mu_2 = \begin{pmatrix} (\kappa - 1 + \beta)y \\ 0 \end{pmatrix} ,$$

and

$$V_2 = \begin{pmatrix} 2\kappa y^2 + 4(\kappa + \beta)(y + 2y^2 + y^3) & 2(\kappa + \beta)(y + y^2) \\ 2(\kappa + \beta)(y + y^2) & (\kappa + \beta)y \end{pmatrix} .$$

Proof Let $f(x) = x^2$ and $g(x) = x$. Define C_n and B_n by the decomposition

$$\sum_{i=1}^p \ell_i^2 = p \int f(x) d(F_n(x) - F^{y_N}(x)) + pF^{y_N}(f) = C_n + pF^{y_N}(f) \, ,$$

$$\sum_{i=1}^p \ell_i = p \int g(x) d(F_n(x) - F^{y_N}(x)) + pF^{y_N}(g) = B_n + pF^{y_N}(g) \, .$$

By the one-sample substitution principle in Theorem 3.16 and the CLT Theorem 3.4, we have

$$\begin{pmatrix} C_n \\ B_n \end{pmatrix} \xrightarrow{\mathscr{D}} N\left(\begin{pmatrix} \mathbb{E} X_f \\ \mathbb{E} X_g \end{pmatrix}, \begin{pmatrix} \mathrm{cov}(X_f, X_f) & \mathrm{cov}(X_f, X_g) \\ \mathrm{cov}(X_g, X_f) & \mathrm{cov}(X_g, X_g) \end{pmatrix} \right).$$

It remains to evaluate the limiting parameters and this results from the following calculations:

$$I_1(f, r) = \frac{h^2}{r^2} \, , \tag{5.52}$$

$$I_1(g,r) = 0\,, \tag{5.53}$$

$$I_2(f) = h^2\,, \tag{5.54}$$

$$I_2(g) = 0\,, \tag{5.55}$$

$$J_1(f,g,r) = \frac{2h^2+2h^4}{r^2}\,, \tag{5.56}$$

$$J_1(f,f,r) = \frac{2h^4+(2h+2h^3)^2 r}{r^3}\,, \tag{5.57}$$

$$J_1(g,g,r) = \frac{h^2}{r^2}\,, \tag{5.58}$$

$$J_2(f,g) = 2h^2+2h^4\,, \tag{5.59}$$

$$J_2(f,f) = (2h+2h^3)^2\,, \tag{5.60}$$

$$J_2(g,g) = h^2\,. \tag{5.61}$$

The results (5.53), (5.55), (5.58) and (5.61) are exactly the same as those found in the proof of Proposition 3.8. The remaining results are found by similar calculations using Proposition 3.6 and their details are omitted. □

Proof (of Theorem 5.12). The result of Lemma 5.13 can be rewritten as:

$$N\begin{pmatrix} p^{-1}\sum_{i=1}^p \ell_i^2 - 1 - \frac{p}{N} - \frac{(\kappa+\beta-1)y}{p} \\ p^{-1}\sum_{i=1}^p \ell_i - 1 \end{pmatrix} \xrightarrow{\mathscr{D}} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{1}{y^2}\cdot V_2\right).$$

Define the function $f(x,y) = \frac{x}{y^2} - 1$, then $U = f(p^{-1}\Sigma_{i=1}^p \ell_i^2,\ p^{-1}\Sigma_{i=1}^p \ell_i)$. We have

$$\begin{aligned}
&\frac{\partial f}{\partial x}(1+\frac{p}{N}+\frac{(\kappa+\beta-1)y}{p},1) = 1\,,\\
&\frac{\partial f}{\partial y}(1+\frac{p}{N}+\frac{(\kappa+\beta-1)y}{p},1) = -2(1+\frac{p}{N}+\frac{(\kappa+\beta-1)y}{p})\,,\\
&f(1+\frac{p}{N}+\frac{(\kappa+\beta-1)y}{p},1) = \frac{p}{N}+\frac{(\kappa+\beta-1)y}{p}\,.
\end{aligned}$$

By the delta method,

$$N\Big(U - f(1+\frac{p}{N}+\frac{(\kappa+\beta-1)y}{p},1)\Big) \xrightarrow{\mathscr{D}} \mathcal{N}(0,\lim C),$$

where

$$C = \begin{pmatrix} \frac{\partial f}{\partial x}(1+\frac{p}{N}+\frac{(\kappa+\beta-1)y}{p},1) \\ \frac{\partial f}{\partial y}(1+\frac{p}{N}+\frac{(\kappa+\beta-1)y}{p},1) \end{pmatrix}^T \cdot\Big(\frac{1}{y^2}V_2\Big)\cdot \begin{pmatrix} \frac{\partial f}{\partial x}(1+\frac{p}{N}+\frac{(\kappa+\beta-1)y}{p},1) \\ \frac{\partial f}{\partial y}(1+\frac{p}{N}+\frac{(\kappa+\beta-1)y}{p},1) \end{pmatrix}$$
$$\longrightarrow 2\kappa\,.$$

Therefore,

$$N(U - \frac{p}{N} - \frac{(\kappa+\beta-1)y}{p}) \xrightarrow{\mathscr{D}} \mathcal{N}(0,2\kappa)\,,$$

that is,

$$NU - p \xrightarrow{\mathscr{D}} \mathcal{N}(\kappa+\beta-1, 2\kappa)\,.$$

The proof of Theorem 5.12 is complete. □

The test based on the asymptotic normal distribution given in Theorem 5.12 is referred as the *corrected John's test (CJ)* for sphericity. A striking fact in this theorem is that as in the normal case, the limiting distribution of CJ is *independent* of the dimension-to-sample-size ratio $y = \lim p/n$. In particular, the limiting distribution derived under classical scheme (p fixed, $n \to \infty$), e.g. the distribution $\frac{2}{p}\chi_f^2 - p$ in the normal case, when used for large p, stays very close to this limiting distribution derived for large-dimensional scheme ($p \to \infty, n \to \infty, p/n \to y \in (0, \infty)$). In this sense, Theorem 5.12 gives a theoretical explanation to the widely observed robustness of John's test against the dimension inflation. Moreover, CJ is also valid for the p larger (or much larger) than n case in contrast to the CLRT where this ratio should be kept smaller than 1 to avoid null eigenvalues.

It is also worth noticing that for real normal data, we have $\kappa = 2$ and $\beta = 0$ so that the theorem above reduces to $NU - p \xrightarrow{\mathscr{D}} \mathcal{N}(1, 4)$. This is exactly the result discussed in Ledoit and Wolf (2002). Besides, if the data has a non-normal distribution but has the same first four moments as the normal distribution, we have again $NU - p \xrightarrow{\mathscr{D}} \mathcal{N}(1, 4)$, which turns out to have a universality property.

Note that the limiting parameters in Theorems 5.11 and 5.12 depend on the parameter β, which is in practice unknown with real data when the 4th order moment of the population does not coincide with the one of a normal population. A consistent estimate of β is thus needed for a practical use of these theorems.

5.7.3 Monte Carlo study

Monte Carlo simulations are conducted to find empirical sizes and powers of CLRT and CJ. In particular, here the following questions are examined: how robust are the tests against non-normal distributed data and what is the range of the dimension to sample ratio p/n where the tests are applicable.

For comparison, the performance of the LW test using the asymptotic $\mathcal{N}(1, 4)$ distribution in (5.50) (Notice however this is the CJ test under normal distribution) and the Chen's test (denoted as C for short) using the asymptotic $\mathcal{N}(0, 4)$ distribution derived in Chen et al. (2010) are evaluated. The nominal test level is set to be $\alpha = 0.05$, and for each pair of (p, n), we run 10000 independent replications.

Consider two scenarios with respect to the random vectors $\mathbf{x}_i$:

(a) $\mathbf{x}_i$ is p-dimensional real random vector from the multivariate normal population $\mathcal{N}(0, \mathbf{I}_p)$. In this case, $\kappa = 2$ and $\beta = 0$.
(b) $\mathbf{x}_i$ consists of iid real random variables with distribution *Gamma*$(4, 2) - 2$ so that x_{ij} satisfies $\mathbb{E}\, x_{ij} = 0$, $\mathbb{E}\, x_{ij}^4 = 4.5$. In this case, $\kappa = 2$ and $\beta = 1.5$.

Table 5.4 reports the sizes of the four tests in these two scenarios for different values of (p, n). When $\{x_{ij}\}$ are normal, LW (=CJ), CLRT and C all have similar empirical sizes tending to the nominal level 0.05 as either p or n increases. But when $\{x_{ij}\}$ are Gamma-distributed, the sizes of LW are higher than 0.1 no matter how large the values of p and n are while the sizes of CLRT and CJ all converge to the nominal level 0.05 as either p or

Table 5.4 *Empirical sizes of LW, CJ, CLRT and C test at 5% significance level based on 10000 independent applications with real* $\mathcal{N}(0,1)$ *random variables and with real Gamma(4,2)-2 random variables.*

(p,n)	$\mathcal{N}(0,1)$			Gamma(4,2)-2			
	LW/CJ	CLRT	C	LW	CLRT	CJ	C
(4,64)	0.0498	0.0553	0.0523	0.1396	0.074	0.0698	0.0717
(8,64)	0.0545	0.061	0.0572	0.1757	0.0721	0.0804	0.078
(16,64)	0.0539	0.0547	0.0577	0.1854	0.0614	0.078	0.0756
(32,64)	0.0558	0.0531	0.0612	0.1943	0.0564	0.0703	0.0682
(48,64)	0.0551	0.0522	0.0602	0.1956	0.0568	0.0685	0.0652
(56,64)	0.0547	0.0505	0.0596	0.1942	0.0549	0.0615	0.0603
(60,64)	0.0523	0.0587	0.0585	0.194	0.0582	0.0615	0.0603
(8,128)	0.0539	0.0546	0.0569	0.1732	0.0701	0.075	0.0754
(16,128)	0.0523	0.0534	0.0548	0.1859	0.0673	0.0724	0.0694
(32,128)	0.051	0.0545	0.0523	0.1951	0.0615	0.0695	0.0693
(64,128)	0.0538	0.0528	0.0552	0.1867	0.0485	0.0603	0.0597
(96,128)	0.055	0.0568	0.0581	0.1892	0.0539	0.0577	0.0579
(112,128)	0.0543	0.0522	0.0591	0.1875	0.0534	0.0591	0.0593
(120,128)	0.0545	0.0541	0.0561	0.1849	0.051	0.0598	0.0596
(16,256)	0.0544	0.055	0.0574	0.1898	0.0694	0.0719	0.0716
(32,256)	0.0534	0.0515	0.0553	0.1865	0.0574	0.0634	0.0614
(64,256)	0.0519	0.0537	0.0522	0.1869	0.0534	0.0598	0.0608
(128,256)	0.0507	0.0505	0.0498	0.1858	0.051	0.0555	0.0552
(192,256)	0.0507	0.054	0.0518	0.1862	0.0464	0.052	0.0535
(224,256)	0.0503	0.0541	0.0516	0.1837	0.0469	0.0541	0.0538
(240,256)	0.0494	0.053	0.0521	0.1831	0.049	0.0533	0.0559
(32,512)	0.0542	0.0543	0.0554	0.1884	0.0571	0.0606	0.059
(64,512)	0.0512	0.0497	0.0513	0.1816	0.0567	0.0579	0.0557
(128,512)	0.0519	0.0567	0.0533	0.1832	0.0491	0.0507	0.0504
(256,512)	0.0491	0.0503	0.0501	0.1801	0.0504	0.0495	0.0492
(384,512)	0.0487	0.0505	0.0499	0.1826	0.051	0.0502	0.0507
(448,512)	0.0496	0.0495	0.0503	0.1881	0.0526	0.0482	0.0485
(480,512)	0.0488	0.0511	0.0505	0.1801	0.0523	0.053	0.0516

n gets larger. This empirically confirms that normal assumptions are needed for the result of Ledoit and Wolf (2002) while the corrected criteria CLRT and CJ (also the C test) do not need such distributional restriction.

As for empirical powers, two alternatives are considered (here, the limiting spectral distributions of $\mathbf{\Sigma}_p$ under these two alternatives differ from that under H_0):

(1) Power 1: $\mathbf{\Sigma}_p$ is diagonal with half of its diagonal elements 0.5 and half 1;
(2) Power 2: $\mathbf{\Sigma}_p$ is diagonal with 1/4 of the elements equal 0.5 and 3/4 equal 1.

Table 5.5 reports the powers of LW(=CJ), CLRT and C when $\{x_{ij}\}$ are distributed as $\mathcal{N}(0,1)$, and of CJ, CLRT and C when $\{x_{ij}\}$ are distributed as Gamma(4,2)-2, for the situation when n equals 64 or 128, with varying values of p and under the above mentioned two alternatives. For $n = 256$ and p varying from 16 to 240, all the tests have powers around 1 under both alternatives so that these values are omitted. And in order to find the trend of these powers, we also present the results when $n = 128$ in Figure 5.1 and Figure 5.2.

Table 5.5 *Empirical powers of LW, CJ, CLRT and C test at* 5% *significance level based on 10000 independent applications with real* $\mathcal{N}(0,1)$ *random variables and with real Gamma(4,2)-2 random variables under two alternatives Power 1 and 2 (see the text for details).*

$\mathcal{N}(0,1)$

(p,n)	Power 1			Power 2		
	LW/CJ	CLRT	C	LW/CJ	CLRT	C
(4,64)	0.7754	0.7919	0.772	0.4694	0.6052	0.4716
(8,64)	0.8662	0.8729	0.8582	0.5313	0.6756	0.5308
(16,64)	0.912	0.9075	0.9029	0.5732	0.6889	0.5671
(32,64)	0.9384	0.8791	0.931	0.5868	0.6238	0.5775
(48,64)	0.9471	0.7767	0.9389	0.6035	0.5036	0.5982
(56,64)	0.949	0.6663	0.9411	0.6025	0.4055	0.5982
(60,64)	0.9501	0.5575	0.941	0.6048	0.3328	0.5989
(8,128)	0.9984	0.9989	0.9986	0.9424	0.9776	0.9391
(16,128)	0.9998	1	0.9998	0.9698	0.9926	0.9676
(32,128)	1	1	1	0.9781	0.9956	0.9747
(64,128)	1	1	1	0.9823	0.9897	0.9788
(96,128)	1	0.9996	1	0.9824	0.9532	0.9804
(112,128)	1	0.9943	1	0.9841	0.881	0.9808
(120,128)	1	0.9746	1	0.9844	0.7953	0.9817

$Gamma(4,2)-2$

(p,n)	Power 1			Power 2		
	CJ	CLRT	C	CJ	CLRT	C
(4,64)	0.6517	0.6826	0.6628	0.3998	0.5188	0.4204
(8,64)	0.7693	0.7916	0.781	0.4757	0.5927	0.4889
(16,64)	0.8464	0.8439	0.846	0.5327	0.633	0.5318
(32,64)	0.9041	0.848	0.9032	0.5805	0.5966	0.5667
(48,64)	0.9245	0.7606	0.924	0.5817	0.4914	0.5804
(56,64)	0.9267	0.6516	0.9247	0.5882	0.4078	0.583
(60,64)	0.9288	0.5547	0.9257	0.5919	0.3372	0.5848
(8,128)	0.9859	0.9875	0.9873	0.8704	0.9294	0.8748
(16,128)	0.999	0.999	0.9987	0.9276	0.9699	0.9311
(32,128)	0.9999	1	0.9999	0.9582	0.9873	0.9587
(64,128)	1	0.9998	1	0.9729	0.984	0.9727
(96,128)	1	0.999	1	0.9771	0.9482	0.9763
(112,128)	1	0.9924	1	0.9781	0.8747	0.9763
(120,128)	1	0.9728	1	0.9786	0.7864	0.977

The behaviour of Power 1 and Power 2 in each figure related to the three statistics are similar, except that Power 1 is much higher compared with Power 2 for a given dimension design (p,n) and any given test for the reason that the first alternative differs more from the null than the second one. The powers of LW (in the normal case), CJ (in the Gamma case) and C are all monotonically increasing in p for a fixed value of n. But for CLRT, when n is fixed, the powers first increase in p and then become decreasing when p is getting close to n. This can be explained by the fact that when p is close to n, some of the eigenvalues of $\mathbf{S}_n$ are getting close to zero, causing the CLRT nearly degenerate and losing power.

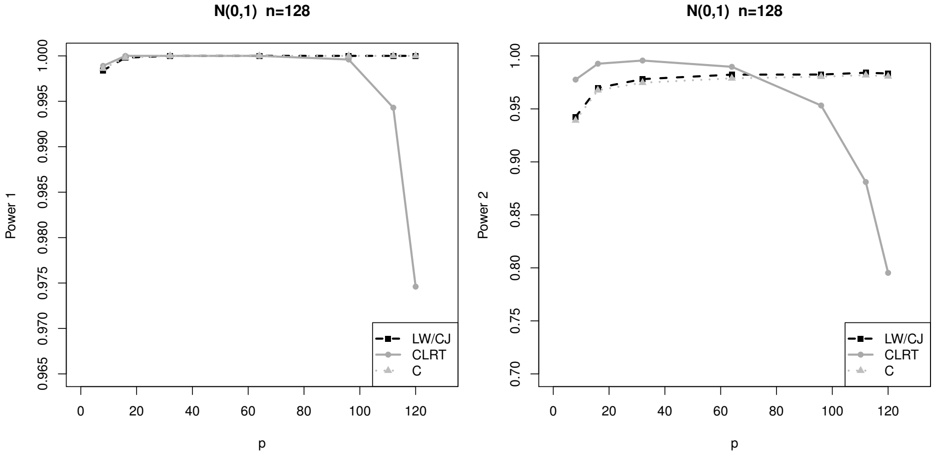


Figure 5.1 Empirical powers of LW/CJ, CLRT and C test at 5% significance level based on 10000 independent applications with real $\mathcal{N}(0,1)$ random variables for fixed $n = 128$ under two alternatives Power 1 and 2 (see the text for details).

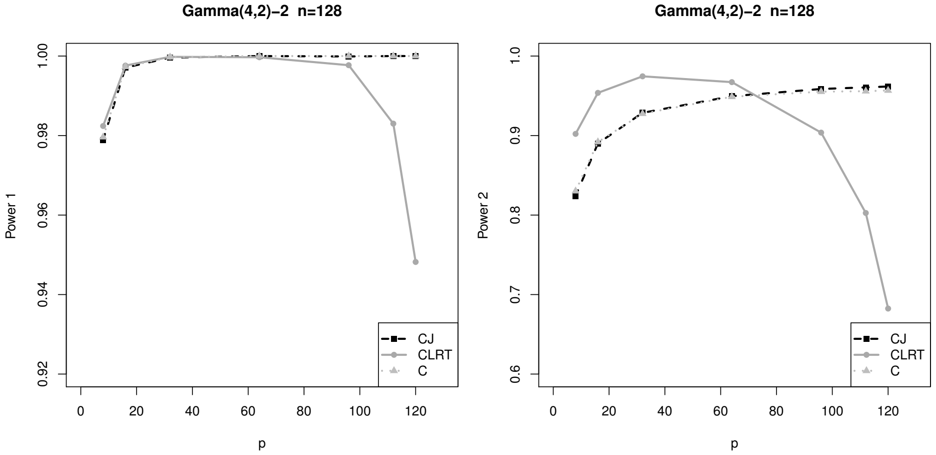


Figure 5.2 Empirical powers of CJ, CLRT and C test at 5% significance level based on 10000 independent applications with real Gamma(4,2)-2 random variables for fixed $n = 128$ under two alternatives Power 1 and 2 (see the text for details).

Besides, in the normal case the trend of C's power is very much alike of those of LW while in the Gamma case it is similar with those of CJ under both alternatives. And in most of the cases (especially in large p case), the power of C test is slightly lower than LW (in the normal case) and CJ (in the Gamma case).

Lastly, consider the performance of CJ and C when p is larger than n. Empirical sizes and powers are presented in Table 5.6. We choose the variables to be distributed as Gamma(4,2)-2 since CJ reduces to LW in the normal case, and Ledoit and Wolf (2002) has already reported the performance of LW when p is larger than n. From the table, we see that when p is larger than n, the size of CJ is still correct and it is always around the nominal level 0.05 as the dimension p increases and the same phenomenon exists for C test.

Table 5.6 *Empirical sizes and powers (Power 1 and 2) of CJ test and C test at* 5% *significance level based on 10000 independent applications with real Gamma(4,2)-2 random variables when* $p \geq n$.

(p, n)	CJ			C		
	Size	Power 1	Power 2	Size	Power 1	Power 2
(64,64)	0.0624	0.9282	0.5897	0.0624	0.9257	0.5821
(320,64)	0.0577	0.9526	0.612	0.0576	0.9472	0.6059
(640,64)	0.0558	0.959	0.6273	0.0562	0.9541	0.6105
(960,64)	0.0543	0.9631	0.6259	0.0551	0.955	0.6153
(1280,64)	0.0555	0.9607	0.6192	0.0577	0.9544	0.6067

Here again power functions are evaluated under the same two alternatives Power 1 and Power 2 as above. The sample size is fixed to $n = 64$ and the ratio p/n varies from 1 to 20. Power 1 is in general much higher than Power 2 for the same reason that the first alternative is easier to be distinguished from H_0. Besides, the powers under both alternatives all increase monotonically for $1 \leq p/n \leq 15$. However, when p/n is getting larger, say $p/n = 20$, we can observe that its size is a little larger and powers a little drop (compared with $p/n = 15$) but overall, it still behaves well, which can be considered as free from the assumption constraint "$p/n \to y$". Besides, the powers of CJ are always slightly higher than those of C in this "large p small n" setting.

Since the asymptotic distribution for the CLRT and CJ are both derived under the "Marcenko-Pasture scheme" (i.e $p/n \to y \in (0, \infty)$), if p/n is getting too large ($p \gg n$), it seems that the limiting results provided in this paper will loose accuracy.

Summarizing all these findings from this Monte-Carlo study, the overall figure is the following: when the ratio p/n is much lower than 1 (say smaller than 1/2), it is preferable to employ CLRT (than CJ, LW or C); while this ratio is higher, CJ (or LW for normal data) becomes more powerful (slightly more powerful than C).

We conclude the section by the following remarks.

(i) The asymptotic distributions derived for the CLRT and the CJ test are universal in the sense that they depend on the distribution of the observations only through its first four moments;

(ii) These corrected test procedures improve quickly when either the dimension p or the sample size n gets large. In particular, for a given sample size n, within a wide range of values of p/n, higher dimensions p lead to better performance of these corrected test statistics.

(iii) CJ is particularly robust against the dimension inflation. Monte-Carlo study shows that for a small sample size $n = 64$, the test is effective for $0 < p/n \leq 20$.

Notes

Large-dimensional tests on covariance matrices developed in §5.6 are due to Bai et al. (2009).

For the sphericity test with large-dimensional data, results in §5.7 are due to Wang and Yao (2013). Related work is done in Ledoit and Wolf (2002), which confirms the robustness of John's test in large-dimensions; however, these results assume a normal population. Following the idea of this paper, Chen et al. (2010) proposed to use a family of well selected U-statistics to test the sphericity; this test is compared by simulation in §5.7.3. Another criterion is proposed in Srivastava et al. (2011) following Srivastava (2005) under non-normality, but with the moment condition $\mathbb{E}|x_{ij}|^4 = 3 + O(p^{-\epsilon})$, for some $\epsilon > 0$ which essentially matches asymptotically the normal case where $\mathbb{E}|x_{ij}|^4 = 3$. It is worth noticing that John's test under normality assumption has been extended for "ultra-dimensional" data, i.e. $p \wedge n \to \infty$ and $p/n \to \infty$ in Birke and Dette (2005).

6

Large-dimensional spiked population models

6.1 Introduction

Principal component analysis is a widely-used data exploration tool in multivariate statistics. Consider a p-variate population $\mathbf{x}$ with population covariance matrix $\mathbf{\Sigma} = \mathrm{cov}(\mathbf{x})$ and let $\mathbf{S}_n$ be the sample covariance matrix based on a sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of size n. In a principal component analysis, one seeks the successively orthogonal directions that maximally explain the variation in the data, that is,

$$\lambda_j = \max \left\{ \mathbf{u}'\mathbf{S}_n\mathbf{u} \ : \ \|\mathbf{u}\| = 1, \ \mathbf{u} \perp \mathbf{u}_1, \ldots, \mathbf{u}_{j-1} \ , \ \ j = 1, \ldots, n \wedge p \right\}.$$

Here a key question emerges: how many principal components should be retained as being “significant”? The *scree plot* is one of the many graphical and informal methods that have been proposed. One plots the ordered sample eigenvalues, and looks for an “elbow”, or other break between presumably significant and presumably unimportant components.

Two such scree plots are given in Figures 6.1 and 6.2. The data shown on Figure 6.1 is a small speech data set collecting 162 instances of a phoneme “dcl” spoken by about 50 males. Each instance is calculated as a periodogram on 256 points. So here $n = 162$ and $p = 256$. The scree-plot shows clearly three large sample eigenvalues, but what about the fourth, fifth, etc.?

The data set on Figure 6.2 consists in daily returns of 488 stock prices listed in the S&P 500 index from September, 2007 to September 2011 (1001 trading days, 12 stocks have been removed because of missing values). So here $p = 488$ and $n = 1000$. Among the 488 eigenvalues of the empirical correlation matrix, the 10 largest are

$$\{237.95, \ 17.76, \ 14.00, \ 8.76, \ 5.29, \ 4.85, \ 4.39, \ 3.49, \ 3.08, \ 2.71\}.$$

These values clearly separate from the rest of 478 eigenvalues and should then be included as principal components. Are there any other valuable principal components among the rest? From the scree-plot shown on the figure, it is again unclear how many other principal components are significant.

In both examples, it appears clearly that the eigenvalues of a sample covariance matrix (or sample correlation matrix) from real-life data can be distinguished in two general area: the *bulk*, which refers to the properties of the majority of eigenvalues, and the *extremes*, which addresses the (first few) largest and smallest eigenvalues. However, as experienced by the elbow rule, the exact cutting-point between these two areas can hardly be known.

The Marčenko-Pastur law and its generalisation introduced in Chapter 2 provide an

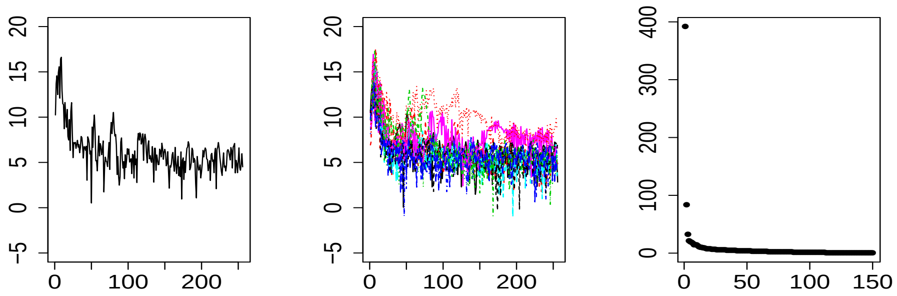


Figure 6.1 (a) a single instance of a periodogram from the phoneme data-set; (b) ten instances, to indicate variability; (c) scree-plot of eigenvalues in phoneme example.

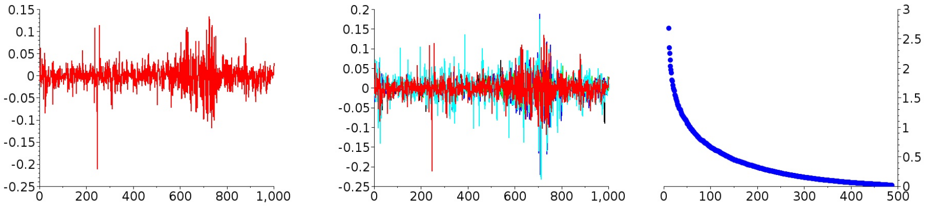


Figure 6.2 (a) a single time series of daily-returns from the S&P 500 data-set; (b) 5 instances, to indicate variability; (c) scree-plot of 478 sample eigenvalues (excluding the 10 largest).

accurate model for the description of the bulk spectrum. It is well established that the bulk spectrum of a sample covariance matrix for a large class of large-dimensional populations converges to a Marčenko-Pastur distribution, see Theorem 2.14.

The situation for extreme eigenvalues is more intricate. In the so-called *null case*, i.e. the population $\mathbf{x}$ has p i.i.d. standardised components (as in Theorem 2.9), we have the following result.

Theorem 6.1 *Let $\{x_{ij}\}$, $i, j = 1, 2, \ldots$, be a double array of i.i.d. complex-valued random variables with mean 0, variance 1 and finite fourth-order moment. Consider the sample covariance matrix $\mathbf{S}_n$ defined in Eq.*(2.7) *where $\mathbf{x}_k = (x_{1k}, \ldots, x_{pk})'$ and denote its eigenvalues in a decreasing order as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. When $p/n \to y > 0$,*

$$\lambda_1 \xrightarrow{a.s.} b_y = (1 + \sqrt{y})^2, \tag{6.1}$$

$$\lambda_{\min} \xrightarrow{a.s.} a_y = (1 - \sqrt{y})^2, \tag{6.2}$$

where

$$\lambda_{\min} = \begin{cases} \lambda_p, & \textit{for } p \leq n, \\ \lambda_{p-n+1}, & \textit{otherwise.} \end{cases}$$

In other words, in the null case, the largest and the smallest eigenvalues are packed together near the right edge b_y and the left edge a_y of the limiting Marčenko-Pastur law, respectively. In addition, the fluctuation of the largest eigenvalue λ_1 has been characterised in this null case. Let

$$\mu_{np} = \frac{1}{n}\left\{(n-1)^{\frac{1}{2}} + p^{\frac{1}{2}}\right\}^2,$$
$$\sigma_{np} = \frac{1}{n}\{(n-1)^{\frac{1}{2}} + p^{\frac{1}{2}}\}\left\{(n-1)^{-\frac{1}{2}} + p^{-\frac{1}{2}}\right\}^{\frac{1}{3}}.$$

Notice that for large p and n, $\mu_{np} \simeq b = (1+\sqrt{y})^2$ (right edge of the Marčenko-Pastur law). Then, under the same conditions as in Theorem 6.1,

$$\frac{\lambda_1 - \mu_{np}}{\sigma_{np}} \xrightarrow{\mathscr{D}} F_1, \tag{6.3}$$

where F_1 is the Tracy-Widom law of order 1 whose distribution function is given by

$$F_1(s) = \exp\left\{\frac{1}{2}\int_s^\infty [q(x) + (x-s)^2 q^2(x)]dx\right\}, \quad s \in \mathbb{R},$$

where q solves the Painlevé II differential equation

$$q''(x) = xq(x) + 2q^3(x).$$

The distribution function F_1 has no closed-form formula and is evaluated by numerical software. In particular, F_1 has mean -1.21 and standard deviation 1.27.

However, the largest sample eigenvalues of the phoneme data set or the S&P 500 daily returns in Figures 6.1 and 6.2 clearly separate from the bulk and they are not packed together as in the null case. These largest eigenvalues do not obey the limiting laws above for the null case. Such new situations are hereafter referred as *non-null cases*.

Nicely enough, much of the behaviour of the extreme eigenvalues in such non null cases can be mathematically explained within the framework of *spiked population model*. In its simplest form, the population covariance matrix $\mathbf{\Sigma}$ in a spiked population model has only m non-unit eigenvalues,

$$\text{spec}\,(\mathbf{\Sigma}) = \{\alpha_1, \ldots, \alpha_m, 1, \ldots, 1\}. \tag{6.4}$$

This model is referred as *Johnstone's spiked population model*. The m non-unit eigenvalues are called *spike eigenvalues*.

Assume further $n \to \infty$, $p/n \to y > 0$. As the number of spikes m is fixed, it is easily seen that the ESD of $\mathbf{S}_n$ still converges to the Marčenko-Pastur law (Theorem 2.9). However, the asymptotic behaviour of the extreme eigenvalues of $\mathbf{S}_n$ as well as the associated eigenvectors is greatly modified by the m spike eigenvalues. This chapter is devoted to a detailed account of such modifications due to the large dimension p.

6.2 Limits of spiked sample eigenvalues

The spiked population model above is extended as follows. Assume that the observed vectors are $\mathbf{x}_i = \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{y}_i$, $i = 1, \ldots, n$ where $\mathbf{y}_i$ are i.i.d. p-dimensional vectors with mean 0, variance 1 and i.i.d. components (this model is already used in §2.4). Therefore, $\{\mathbf{x}_i\}$ is a sequence of i.i.d. random vectors with mean $\mathbf{0}$ and population covariance matrix $\mathbf{\Sigma}$.

Furthermore, $\mathbf{\Sigma}$ has the structure:

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_p \end{pmatrix}. \tag{6.5}$$

It is assumed that

(i) $\mathbf{\Lambda}$ is of size $m \times m$ where m is a fixed integer. The eigenvalues of $\mathbf{\Lambda}$ are $\alpha_1 > \cdots > \alpha_K > 0$ of respective multiplicity $m_1, \ldots, m_K$ (so that $m = m_1 + \cdots + m_K$). Denote by J_j the set of m_j indexes of α_j in the matrix $\mathbf{\Sigma}$.

(ii) The ESD H_p of $\mathbf{V}_p$ converges to a nonrandom limiting distribution H.

(iii) The sequence of the largest eigenvalue of $\mathbf{\Sigma}$ is bounded.

(iv) The eigenvalue $\{\beta_{pj}\}$ of $\mathbf{V}_p$ are such that

$$\sup_j d(\beta_{pj}, \Gamma_H) = \varepsilon_p \to 0,$$

where $d(x, A)$ denotes the distance of x to a set A and Γ_H the support of H.

Definition 6.2 An eigenvalue α of $\mathbf{\Lambda}$ is called a *generalised spike*, or simply a *spike*, if $\alpha \notin \Gamma_H$.

Such a model is called a *generalised spiked population model*. Roughly speaking, the population eigenvalues of $\mathbf{\Sigma}$ are composed of a main spectrum made with the $\{\beta_{pj}\}$'s, and a small and finite spectrum of m spikes eigenvalues that are well separated from the main spectrum (in the sense of Definition 6.2).

The following assumptions are also needed.

(v) $\mathbb{E}y_{ij} = 0$, $\mathbb{E}|y_{ij}|^2 = 1$ and $\mathbb{E}|y_{ij}|^4 < \infty$.

(vi) $p/n \to y > 0,\ p \wedge n \to \infty$.

The analysis below is carried out using a decomposition into blocks of size m and $p-m$, respectively:

$$\mathbf{x}_i = \begin{pmatrix} \mathbf{x}_{1i} \\ \mathbf{x}_{2i} \end{pmatrix}, \quad \mathbf{y}_i = \begin{pmatrix} \mathbf{y}_{1i} \\ \mathbf{y}_{2i} \end{pmatrix}.$$

Define the sample covariance matrix as

$$\mathbf{S}_n = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k \mathbf{x}_k^* = \frac{1}{n}\begin{pmatrix} \mathbf{X}_1\mathbf{X}_1^* & \mathbf{X}_1\mathbf{X}_2^* \\ \mathbf{X}_2\mathbf{X}_1^* & \mathbf{X}_2\mathbf{X}_2^* \end{pmatrix} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix}, \tag{6.6}$$

where

$$\mathbf{X}_1 = (\mathbf{x}_{11}, \cdots, \mathbf{x}_{1n}), \quad \mathbf{X}_2 = (\mathbf{x}_{21}, \cdots, \mathbf{x}_{2n}). \tag{6.7}$$

Define also the analogous decomposition for the $\mathbf{y}_i$ vectors to the data matrices $\mathbf{Y}_1$ and $\mathbf{Y}_2$ satisfying

$$\mathbf{X}_1 = \mathbf{\Lambda}^{\frac{1}{2}}\mathbf{Y}_1, \quad \mathbf{X}_2 = \mathbf{V}_p^{\frac{1}{2}}\mathbf{Y}_2. \tag{6.8}$$

An eigenvalue λ_i of $\mathbf{S}_n$ that is not an eigenvalue of $\mathbf{S}_{22}$ satisfies

$$0 = |\lambda_i \mathbf{I}_p - \mathbf{S}_n| = |\lambda_i \mathbf{I}_{p-m} - \mathbf{S}_{22}| \cdot |\lambda_i \mathbf{I}_m - \mathbf{K}_n(\lambda_i)|, \tag{6.9}$$

where

$$\mathbf{K}_n(l) = \mathbf{S}_{11} + \mathbf{S}_{12}(l\mathbf{I}_{p-m} - \mathbf{S}_{22})^{-1}\mathbf{S}_{21}.$$

Since for large n, it will holds eventually that $|\lambda_i \mathbf{I}_{p-m} - \mathbf{S}_{22}| \neq 0$,

$$|\lambda_i \mathbf{I}_m - \mathbf{K}_n(\lambda_i)| = 0. \tag{6.10}$$

Consider now a real number l outside the support of the LSD of $\mathbf{S}_{22}$ and the goal is to find the limit of the random matrix $\mathbf{K}_n(l)$ with fixed dimension m. It holds

$$\begin{aligned}\mathbf{K}_n(l) &= \mathbf{S}_{11} + \mathbf{S}_{12}(l\mathbf{I}_{p-m} - \mathbf{S}_{22})^{-1}\mathbf{S}_{21} \\ &= \frac{1}{n}\mathbf{X}_1 \left\{ \mathbf{I}_n + \frac{1}{n}\mathbf{X}_2^*(l\mathbf{I}_{p-m} - \frac{1}{n}\mathbf{X}_2\mathbf{X}_2^*)^{-1}\mathbf{X}_2 \right\} \mathbf{X}_1^* \\ &= \frac{l}{n}\mathbf{X}_1(l\mathbf{I}_n - \frac{1}{n}\mathbf{X}_2^*\mathbf{X}_2)^{-1}\mathbf{X}_1^* \\ &= \frac{l}{n}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{Y}_1(l\mathbf{I}_n - \frac{1}{n}\mathbf{X}_2^*\mathbf{X}_2)^{-1}\mathbf{Y}_1^*\mathbf{\Lambda}^{\frac{1}{2}}. \end{aligned} \tag{6.11}$$

The above deduction has used the following identity: for $l \neq 0$ which is not an eigenvalue of $\mathbf{A}^*\mathbf{A}$,

$$\mathbf{I}_n + \mathbf{A}(l\mathbf{I}_{p-m} - \mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^* \equiv l(l\mathbf{I}_n - \mathbf{A}\mathbf{A}^*)^{-1}. \tag{6.12}$$

Since by assumption l is outside the support of the LSD $F_{y,H}$ of $\mathbf{S}_{22}$, for large enough n, the operator norm of $(l\mathbf{I}_n - \frac{1}{n}\mathbf{X}_2^*\mathbf{X}_2)^{-1}$ is bounded. By the law of large numbers,

$$\begin{aligned}\mathbf{K}_n(l) &= \mathbf{\Lambda} \cdot \left[l \cdot \frac{1}{n}\mathrm{tr}\left(l\mathbf{I}_n - \frac{1}{n}\mathbf{X}_2^*\mathbf{X}_2 \right)^{-1} \right] + o_{a.s.}(1) \\ &= -\mathbf{\Lambda} \cdot l\underline{s}(l) + o_{a.s.}(1), \end{aligned} \tag{6.13}$$

where $\underline{s}$ is the Stieltjes transform of the LSD of $\frac{1}{n}\mathbf{X}_2^*\mathbf{X}_2$.

If for some sub-sequence $\{i\}$ of $\{1, \ldots, n\}$, $\lambda_i \to l$ a.s. where l is as above, it holds then $\mathbf{K}_n(\lambda_i) \to -\mathbf{\Lambda} l\underline{s}(l)$ a.s. Therefore necessarily, l is an eigenvalue of $-\mathbf{\Lambda} l\underline{s}(l)$, that is $l = -\alpha_j l\underline{s}(l)$, or equivalently

$$\underline{s}(l) = -1/\alpha_j. \tag{6.14}$$

Recall the ψ-function defined in (2.20),

$$\psi(\alpha) = \psi_{y,H}(\alpha) = \alpha + y\int \frac{t\alpha}{\alpha - t}dH(t), \tag{6.15}$$

which is the functional inverse of the function $x \mapsto -1/\underline{s}(x)$. Moreover, ψ is well-defined for all $\alpha \notin \Gamma_H$. Overall, we have proved that if such limit l exists, l necessarily satisfies the equation

$$l = \psi(\alpha_j) \ ,$$

for some α_j. Furthermore, by Proposition 2.17, $l = \psi(\alpha_j)$ is outside the support of the LSD $F_{y,H}$ if and only if $\psi'(\alpha_j) > 0$.

In summary, the above analysis shows that if α_j is a spike eigenvalue such that $l = \psi(\alpha_j)$ is the limit for some sub-sequence of sample eigenvalues $\{\lambda_i\}$, then necessarily $\psi'(\alpha_j) > 0$. It turns out that this is also a sufficient condition for such limit to exist as given in the following theorem.

Theorem 6.3 *(a) For a spike eigenvalue α_j satisfying*

$$\psi'(\alpha_j) > 0,$$

there are m_j sample eigenvalues λ_i of $\mathbf{S}$ *with $i \in J_j$ such that*

$$\lambda_i \xrightarrow{a.s.} \psi_j = \psi(\alpha_j), \tag{6.16}$$

(b) *For a spike eigenvalue α_j satisfying*

$$\psi'(\alpha_j) \le 0,$$

there are m_j sample eigenvalues λ_i of $\mathbf{S}$ *with $i \in J_j$ such that*

$$\lambda_i \xrightarrow{a.s.} \gamma_j,$$

where γ_j is the γ-th quantile of $F_{y,H}$ with $\gamma = H(-\infty, \alpha_j]$ and H the LSD of $\mathbf{V}_p$*.*

For the proof of this theorem, interested reader can refer to the references given at the end of the chapter. The theorem separates spike eigenvalues into two groups; those with a positive ψ' can be identified as *fundamental spikes* and the others with a non-positive ψ' as *non fundamental spikes*. A fundamental spike α_j is that for large enough n, exactly m_j sample eigenvalues will cluster in a neighbourhood of $\psi_{y,H}(\alpha_j)$ which is outside the support of the LSD $F_{y,H}$. These limits are also seen as *outliers* compared to the bulk spectrum of the sample covariance matrix. Sample eigenvalues converging to a limit $\psi(\alpha_j)$ lying outside the support of the LSD are hereafter referred as *spiked sample eigenvalues*.

Notice that the separation above between fundamental and non-fundamental spike eigenvalues depend not only on the base population spectral distribution H but also on the limiting ratio y. For instance, it can be seen from Eq. (6.15) that for fixed H and when $y \to 0$, the function $\psi_{y,H}$ tends to the identity map so that ψ' tends to the constant function 1. This means that provided y is small enough, any spike eigenvalue α_j is a fundamental spike and there will be m_j spiked sample eigenvalues converging to $\psi_{y,H}(\alpha_j)$. In particular, when y is close to zero (i.e. p is much smaller than n), we will have $\psi_{y,H}(\alpha_j) \simeq \alpha_j$. In other words, this scheme recovers the consistency property usually observed under a low-dimensional scenario, namely the sample eigenvalues converge all to the corresponding population eigenvalues when the sample size is much larger than the dimension p.

For the function ψ in Eq. (6.15), we have

$$\psi'(\alpha) = 1 - y \int \frac{t^2}{(\alpha - t)^2} dH(t)\,, \qquad \psi''(\alpha) = 2y \int \frac{t^2}{(\alpha - t)^3} dH(t)\,.$$

Assume that H has a compact support, $\Gamma_H = [\theta, \omega]$ with edge points $\theta \le \omega$. From the expressions of the derivatives, it is easily seen that

(i) for $\alpha < \theta$, ψ is concave and varies from $-\infty$ to $-\infty$ where $\psi' = 0$ at a unique point, say ζ_1. Therefore, any spike $\alpha < \zeta_1$ is a fundamental spike, while a spike $\zeta_1 \le \alpha < \theta$ is a non-fundamental one.

(ii) for $\alpha > \omega$, ψ is convex and varies from ∞ to ∞ where $\psi' = 0$ at a unique point, say ζ_2. Therefore, any spike $\alpha > \zeta_2$ is a fundamental spike, while a spike $\omega < \alpha \le \zeta_2$ is a non-fundamental one.

6.2.1 Johnstone's spiked population model

For Johnstone's spiked population model (6.4), $\mathbf{V}_p = \mathbf{I}_{p-m}$ and PSD $H = \delta_1$. We have

$$\psi(\alpha) = \alpha + \frac{y\alpha}{\alpha - 1},$$

and

$$\psi'(\alpha) = 1 - \frac{y}{(\alpha - 1)^2}.$$

This particular ψ-function is plotted on Figure 6.3 for $y = \frac{1}{2}$ and has the following properties,

- its range equals $(-\infty, a_y] \cup [b_y, \infty)$;
- $\psi(1 - \sqrt{y}) = a_y$, $\psi(1 + \sqrt{y}) = b_y$;
- $\psi'(\alpha) > 0 \Leftrightarrow |\alpha - 1| > \sqrt{y}$, i.e. $\zeta_{1,2} = 1 \pm \sqrt{y}$.

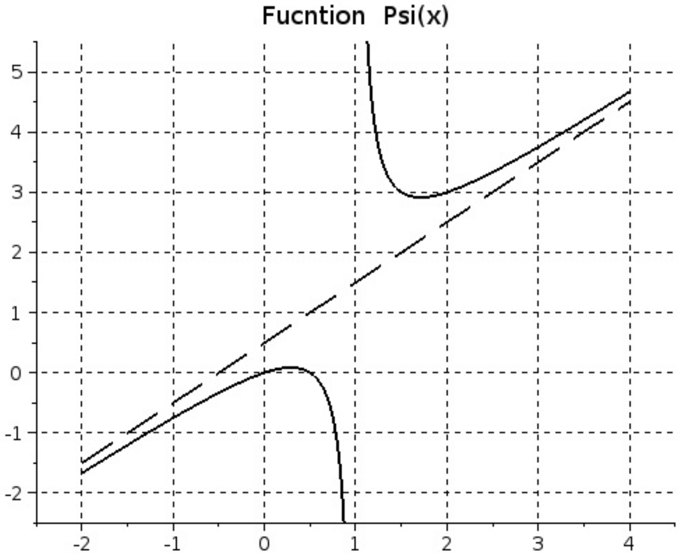


Figure 6.3 The function $\alpha \mapsto \psi(\alpha) = \alpha + y\alpha/(\alpha - 1)$ which maps a spike eigenvalue α to the limit of an associated sample eigenvalue in Johnstone's spiked population model. Figure with $y = \frac{1}{2}$; $[1 \mp \sqrt{y}] = [0.293,\ 1.707]$; $[(1 \mp \sqrt{y})^2] = [0.086,\ 2.914]$.

The exact content of Theorem 6.3 for Johnstone's spiked population model is summarised in the following corollary.

Corollary 6.4 *When* $\mathbf{V}_p = \mathbf{I}_{p-m}$, *it holds that*

(i) *large fundamental spikes: for* $\alpha_j > 1 + \sqrt{y}$,

$$\lambda_i \xrightarrow{a.s.} \alpha_j + \frac{y\alpha_j}{\alpha_j - 1}, \quad i \in J_j \ ;$$

(ii) *large non-fundamental spikes: for* $1 < \alpha_j \leq 1 + \sqrt{y}$,

$$\lambda_i \xrightarrow{a.s.} (1 + \sqrt{y})^2; \quad i \in J_j \ ;$$

(iii) *small non-fundamental spikes: for* $1 - \sqrt{y} \le \alpha_j <$ *with* $y < 1$, *or* $\alpha_j < 1$ *with* $y \ge 1$,

$$\lambda_i \xrightarrow{a.s.} (1 - \sqrt{y})^2, \quad i \in J_j\,.$$

(iv) *small fundamental spikes: for* $\alpha_j < 1 - \sqrt{y}$ *with* $y < 1$,

$$\lambda_i \xrightarrow{a.s.} \alpha_j + \frac{y\alpha_j}{\alpha_j - 1}, \quad i \in J_j\,.$$

It is worth noticing that when $y \ge 1$, a fundamental spike is necessarily greater than 1 as $\alpha < 1 - \sqrt{y}$ becomes impossible.

6.2.2 An example with non-extreme spike eigenvalues

When Γ_H has several compact components, say $\Gamma_H = \cup_{1\le j\le K}[\theta_j, \omega_j]$, the situation for extreme spike eigenvalues on the left of θ_1 or on the right of ω_K is similar to previously, i.e.

(i) for $\alpha < \theta_1$, ψ is concave and varies from $-\infty$ to $-\infty$ where $\psi' = 0$ at a unique point, say ζ_1. Therefore, any spike $\alpha < \zeta_1$ is a fundamental spike, while a spike $\zeta_1 \le \alpha < \theta$ is a non-fundamental one.

(ii) for $\alpha > \omega_K$, ψ is convex and varies from ∞ to ∞ where $\psi' = 0$ at a unique point, say ζ_2. Therefore, any spike $\alpha > \zeta_2$ is a fundamental spike, while a spike $\omega < \alpha \le \zeta_2$ is a non-fundamental one.

However, for *non-extreme spike eigenvalues*, i.e. spikes lying between the K support intervals $[\theta_j, \omega_j]$, the situation is a bit more complicate. Such a ψ-function is given in Example 2.18 where $\psi = \psi_{0.3,H}$ with $H = \frac{1}{3}(\delta_1 + \delta_4 + \delta_{10})$, see also Figure 2.3 that depicts its use for the determination of the support of the corresponding LSD $F_{0.3,H}$; this support consists in two intervals [0.32, 1.37] and [1.67, 18.00].

Consider next a spiked covariance matrix $\mathbf{\Sigma}$ for which the LSD remains the same $F_{0.3,H}$. Precisely, $\mathbf{\Sigma}$ is diagonal with three base eigenvalues $\{1, 4, 10\}$, nearly $p/3$ times for each of them, and there are four spike eigenvalues $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (15,\ 6,\ 2,\ 0.5)$, with respective multiplicities $(m_k) = (3, 2, 2, 2)$ so that $m = \sum m_k = 9$. The limiting population-sample ratio is taken to be $y = 0.3$ while the population spectral distribution H is the uniform distribution on $\{1, 4, 10\}$. For simulation, we use $p - m = 600$ so that $\mathbf{\Sigma}$ has the following 609 eigenvalues:

$$\text{spec}\,(\mathbf{\Sigma}) = \{\ 15,\ 15,\ 15,\ \underbrace{10, \ldots, 10}_{200},\ 6,\ 6,\ \underbrace{4, \ldots, 4}_{200},\ 2,\ 2, \underbrace{1, \ldots, 1}_{200},\ 0.5,\ 0.5\ \}.$$

From the table

spike α_k	15	6	2	0.5
multiplicity n_k	3	2	2	2
$\psi'(\alpha_k)$	$+$	$-$	$+$	$+$
$\psi(\alpha_k)$	18.65	5.82	1.55	0.29
descending ranks	1, 2, 3	204, 205	406, 407	608, 609

we see that 6 is a non-fundamental spike for (y, H) while the three others are fundamental ones. By Theorem 6.3, we know that

- the 7 spiked sample eigenvalues $\lambda_j^{\mathbf{S}_n}$ with $j \in \{1,\ 2,\ 3,\ 406,\ 407,\ 608,\ 609\}$ associated to fundamental spikes tend to 18.65, 1.55 and 0.29, respectively, which are located outside the support of limiting distribution $F_{0.3,H}$;
- the two sample eigenvalues $\lambda_j^{\mathbf{S}_n}$ with $j = 204, 205$ associated to the non-fundamental spike 6 tend to a limit located inside the support, the γ-th quantile of the limiting distribution G where $\gamma = H(0, 6) = 2/3$.

These facts are illustrated by a simulated sample displayed in Figure 6.4.

6.3 Limits of spiked sample eigenvectors

Theorem 6.3 can also be used to find the limits of the eigenvectors associated to the spike eigenvalues.

Theorem 6.5 *Let α_j be a spike with $\psi'(\alpha_j) > 0$. For any (normalised) eigenvector $\mathbf{u}_i$ of $\mathbf{S}_n$ associated to a spiked sample eigenvalue λ_i converging to $\psi(\alpha_j)$, define the block decomposition $\mathbf{u}_i = (\mathbf{u}_{1i}', \mathbf{u}_{2i}')'$ with block lengths m and $p - m$, respectively. Then $\mathbf{u}_{1i}$ converges a.s. to an eigenvector of $\mathbf{\Lambda}$ associated to α_j and*

$$\lim \|\mathbf{u}_{1i}\| = d_j := \sqrt{\frac{\alpha_j \psi'(\alpha_j)}{\psi(\alpha_j)}}\ .$$

In consequence,

$$\lim \|\mathbf{u}_{2i}\| = \sqrt{1 - d_j^2}\ ,$$

and if the spike α_j is simple (i.e. $m_j = 1$), the above limiting vector is unique (up to the sign).

Proof Suppose $\lambda_i \to \psi_j$ and $\mathbf{u}_i$ is an eigenvector of $\mathbf{S}_n$ associated to λ_i. By definition,

$$\begin{pmatrix} \lambda_i \mathbf{I}_m - \mathbf{S}_{11} & -\mathbf{S}_{12} \\ -\mathbf{S}_{21} & \lambda_i \mathbf{I}_{p-m} - \mathbf{S}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{u}_{1i} \\ \mathbf{u}_{2i} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}.$$

Therefore,

$$\begin{aligned} (\lambda_i \mathbf{I}_m - \mathbf{S}_{11})\mathbf{u}_{1i} - \mathbf{S}_{12}\mathbf{u}_{2i} &= \mathbf{0} \\ -\mathbf{S}_{21}\mathbf{u}_{1i} + (\lambda_i \mathbf{I}_{p-m} - \mathbf{S}_{22})\mathbf{u}_{2i} &= \mathbf{0}. \end{aligned}$$

Consequently,

$$\mathbf{u}_{2i} = (\lambda_i \mathbf{I}_{p-m} - \mathbf{S}_{22})^{-1}\mathbf{S}_{21}\mathbf{u}_{1i} \tag{6.17}$$

$$(\lambda_i \mathbf{I}_m - \mathbf{K}_n(\lambda_i))\mathbf{u}_{1i} = \mathbf{0}. \tag{6.18}$$

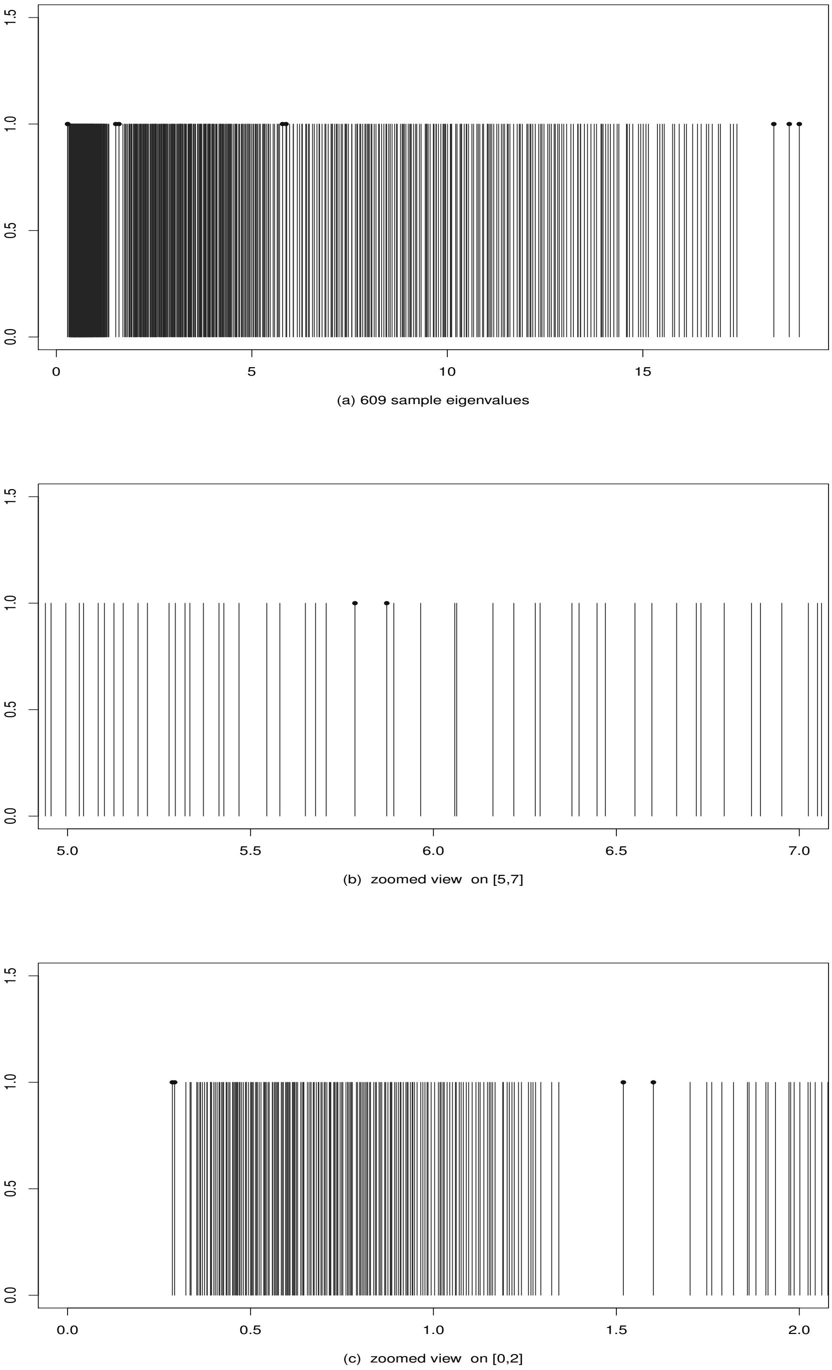
Using (6.13), (6.14) and (6.18),

$$(\mathbf{I}_m - \alpha_j^{-1}\mathbf{\Lambda})\mathbf{u}_{1i} = o_{\text{a.s.}}(1). \tag{6.19}$$

This means that the projection of $\mathbf{u}_{1i}$ onto the orthogonal complement of the eigenspace of $\mathbf{\Lambda}$ associated to α_j tends to 0.

For the limit of $\|\mathbf{u}_{1i}\|$, we have by (6.17),

$$\mathbf{u}_{2i}'\mathbf{u}_{2i} = \mathbf{u}_{1i}'\mathbf{S}_{12}(\lambda_i \mathbf{I}_{p-m} - \mathbf{S}_{22})^{-2}\mathbf{S}_{21}\mathbf{u}_{1i}\ .$$



(a) 609 sample eigenvalues

(b) zoomed view on [5,7]

(c) zoomed view on [0,2]

Figure 6.4 An example of $p = 609$ sample eigenvalues (a), and two zoomed views (b) and (c) on [5,7] and [0,2] respectively. The limiting distribution of the ESD has support $[0.32, 1.37] \cup [1.67, 18.00]$. The 9 sample eigenvalues $\{\lambda_j^{\mathbf{S}_n},\ j = 1, 2, 3, 204, 205, 406, 407, 608, 609\ \}$ associated to the spikes are marked with a blue dot. The eigenvalues $\lambda_{204}^{\mathbf{S}_n}$ and $\lambda_{205}^{\mathbf{S}_n}$ are not spiked and fall inside the support of the LSD, see (b), while the other seven are spiked sample eigenvalues. Gaussian entries are used.

By a method similar to the one leading to (6.13), it holds

$$\mathbf{S}_{12}(\lambda_i \mathbf{I}_{p-m} - \mathbf{S}_{22})^{-2}\mathbf{S}_{21} = \mathbf{\Lambda} \cdot (\underline{s}(\psi_j) + \psi_j \underline{s}'(\psi_j)) + o_{\text{a.s.}}(1).$$

On the other hand, by the definition of the ψ-function $\lambda = \psi(\alpha)$, we have $\alpha\underline{s}(\lambda) = -1$ whenever $\psi'(\alpha) > 0$. Differentiation with respect to α yields

$$\underline{s}'(\lambda) = -\frac{\underline{s}(\lambda)}{\alpha\psi'(\alpha)} = \frac{1}{\alpha^2\psi'(\alpha)} . \tag{6.20}$$

Therefore,

$$\begin{aligned} 1 = \|\mathbf{u}_{1i}\|^2 + \|\mathbf{u}_{2i}\|^2 &= \mathbf{u}_{1i}^*\mathbf{u}_{1i}\left[1 + \alpha_j\{\underline{s}(\psi_j) + \psi_j\underline{s}'(\psi_j)\} + o_{\text{a.s.}}(1)\right] \\ &= \|\mathbf{u}_{1i}\|^2\left\{\frac{\psi_j}{\alpha_j\psi'(\alpha_j)} + o_{\text{a.s.}}(1)\right\} \end{aligned}$$

The first assertion is then proved. The others are obvious. □

Application to the special case of Johnstone's spiked population model with $\mathbf{V}_p = \mathbf{I}_{p-m}$ yields the following

Corollary 6.6 *For Johnstone's spiked population model with* $\mathbf{V}_p = \mathbf{I}_{p-m}$ *it holds that*

(i) *For* $\alpha_j > 1 + \sqrt{y}$, *denote the sample eigenvectors associated to* λ_i *by* $\mathbf{u}_i = (\mathbf{u}_{1i}', \mathbf{u}_{2i}')'$, $i = m_1 + \ldots + m_{j-1} + 1, \cdots, m_1 + \ldots + m_{j-1} + m_j$. *Then* $\mathbf{u}_{1i}$ *tends a.s. to an eigenvector of* $\mathbf{\Sigma}$ *associated to* α_j *with length*

$$d_j = \sqrt{\frac{(\alpha_j - 1)^2 - y}{(\alpha_j - 1)(\alpha_j - 1 + y)}} .$$

Moreover, the length of $\mathbf{u}_{2i}$ *tends to* $\sqrt{1 - d_j^2}$.

(ii) *For* $\alpha_j < 1 - \sqrt{y}$ *and* $y < 1$, *denote the sample eigenvectors associated to* λ_i *by* $\mathbf{u}_i = (\mathbf{u}_{1i}', \mathbf{u}_{2i}')'$, $i = p - m_{j+1} - \ldots - m_k + 1, \cdots, p - m_j - \ldots - m_k$. *Then the same conclusions hold for* $\mathbf{u}_{1i}$ *and the length of* $\mathbf{u}_{2i}$.

Again it is important to note that when $y \geq 1$, a fundamental spike is necessarily greater than 1 so that no question is raised for spike eigenvectors when $\alpha_j < 1 - \sqrt{y}$.

6.4 Central limit theorem for spiked sample eigenvalues

Point-wise limits of spiked sample eigenvalues are derived in Theorem 6.3. In this section, we derive the corresponding central limit theorems.

The m-dimensional random matrix $\mathbf{K}_n(l)$ is introduced in Equation (6.9). From (6.11), we have

$$\begin{aligned} \mathbf{K}_n(l) &= \frac{l}{n}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{Y}_1(l\mathbf{I}_n - \frac{1}{n}\mathbf{Y}_2^*\mathbf{V}_p\mathbf{Y}_2)^{-1}\mathbf{Y}_1^*\mathbf{\Lambda}^{\frac{1}{2}} \\ &= \frac{l}{n}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{Y}_1(l\mathbf{I}_n - \frac{1}{n}\mathbf{Y}_2^*\mathbf{V}_p\mathbf{Y}_2)^{-1}\mathbf{Y}_1^*\mathbf{\Lambda}^{\frac{1}{2}} - \frac{l}{n}\mathbf{\Lambda}\text{tr}(l\mathbf{I}_n - \frac{1}{n}\mathbf{Y}_2^*\mathbf{V}_p\mathbf{Y}_2)^{-1} \\ &\quad -l\mathbf{\Lambda}\left\{\underline{s}_n(l) - \underline{s}(l)\right\} - l\underline{s}(l)\mathbf{\Lambda} \end{aligned}$$

$$:= \frac{l}{\sqrt{n}} \mathbf{\Lambda}^{\frac{1}{2}} \mathbf{R}_n(l) \mathbf{\Lambda}^{\frac{1}{2}} - l\underline{s}(l)\mathbf{\Lambda} + O_p(n^{-1}), \tag{6.21}$$

where $\underline{s}_n$ is the Stieltjes transformation of $\frac{1}{n}\mathbf{Y}_2^*\mathbf{V}_p\mathbf{Y}_2$ and

$$\mathbf{R}_n(l) = \frac{1}{\sqrt{n}}\left(\mathbf{Y}_1(l\mathbf{I}_n - \frac{1}{n}\mathbf{Y}_2^*\mathbf{V}_p\mathbf{Y}_2)^{-1}\mathbf{Y}_1^* - \mathbf{I}_m \mathrm{tr}(l\mathbf{I}_n - \frac{1}{n}\mathbf{Y}_2^*\mathbf{V}_p\mathbf{Y}_2)^{-1}\right) \tag{6.22}$$

is a $m \times m$ random matrix. In the above derivation, we used $\underline{s}_n(l) - \underline{s}(l) = O_p(n^{-1})$ which is a simple consequence of the CLT for linear spectral statistics, e.g. see Theorem 3.9.

6.4.1 Convergence of the matrix-valued process $\{\mathbf{R}_n(l)\}$

For the establishment of a CLT for spiked sample eigenvalues, the key theory is to prove the convergence of the sequence of matrix-valued processes

$$\{\mathbf{R}_n(l)\}, \quad l \in \mathcal{U} ,$$

where $\mathcal{U}$ is a compact set of indexes outside the support of the LSD $F_{y,H}$. The whole proof is quite technical and we describe in this section its main steps and the tools used. For missing details, reader is referred to the original paper Bai and Yao (2008).

The proof consists in two steps.

(i) Establish the tightness of the sequence of processes;
(ii) Establish the finite-dimensional convergence of the process.

For Step (i), it is sufficient to prove the following Lipschitz condition: there exists a constant M such that for any $l_1,\ l_2 \in \mathcal{U}$,

$$\frac{\mathbb{E}|R_n(l_2) - R_n(l_1)|^2}{|l_2 - l_1|^2} \leq M < \infty.$$

For a reference, see e.g. Theorem 12.3 in Billingsley (1968). This Lipschitz condition can be proved using standard calculations of the moments and this part is skipped here.

Step (ii) is developed in the remaining of the section. Consider for any L index values $\{l_j\}$, the distribution of

$$(\mathbf{R}_n(l_1), \ldots, \mathbf{R}_n(l_L)) \tag{6.23}$$

conditioning on $\mathbf{Y}_2$ is derived using Lemma 6.7 below, Next, as it will be seen, this conditional limiting distribution is in fact independent of the conditioning $\mathbf{Y}_2$; it thus equals the unconditional limiting distribution.

This is done in Theorem 6.10 that is based on the Lemma below on CLT for random sesquilinear forms.

Consider a sequence $\{(\mathbf{x}_i, \mathbf{y}_i)_{i\in N}\}$ of i.i.d. complex-valued, zero-mean random vectors belonging to $\mathbb{C}^K \times \mathbb{C}^K$ with finite 4th-moment. We write

$$\mathbf{x}_i = (x_{\ell i}) = \begin{pmatrix} x_{1i} \\ \vdots \\ x_{Ki} \end{pmatrix}, \quad \mathbf{X}_\ell = (x_{\ell 1}, \ldots, x_{\ell n})', \quad 1 \leq \ell \leq K , \tag{6.24}$$

with a similar definition for the vectors $\{\mathbf{Y}_\ell\}_{1\leq\ell\leq K}$. Set $\rho(\ell) = \mathbb{E}[\bar{x}_{\ell 1} y_{\ell 1}]$.

Lemma 6.7 *Let* $\{\mathbf{A}_n = [a_{ij}(n)]\}_n$ *be a sequence of* $n\times n$ *Hermitian matrices with bounded spectral norm and the vectors* $\{\mathbf{X}_\ell, \mathbf{Y}_\ell\}_{1\le\ell\le K}$ *are as defined in* (6.24). *Assume that the following limit exist:*

(a) $\omega = \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} a_{kk}^2(n)$*;*
(b) $\theta = \lim_{n\to\infty} \frac{1}{n} tr\mathbf{A}_n^2 = \lim \frac{1}{n}\sum_{k,j=1}^{n} |a_{k,j}(n)|^2$*; and*
(c) $\tau = \lim_{n\to\infty} \frac{1}{n} tr\mathbf{A}_n\mathbf{A}_n' = \lim \frac{1}{n}\sum_{k,j=1}^{n} a_{k,j}^2(n)$.

Then the K-dimensional complex random vector

$$\mathbf{Z}_n = (Z_{n\ell})_{\ell=1}^K, \quad Z_{n\ell} = \frac{1}{\sqrt{n}}(\mathbf{X}_\ell^*\mathbf{A}_n\mathbf{Y}_\ell - \rho_\ell tr\mathbf{A}_n),$$

converges to a zero-mean, complex normal random vector $\mathbf{Z}$ *(the joint distribution of real part and imaginary part follows a* $2K$*-dimensional real normal distribution) whose moment generating function equals*

$$f(\mathbf{c}) = \mathbb{E}e^{\mathbf{c}'\mathbf{Z}} = \exp(\frac{1}{2}\mathbf{c}'\mathbf{B}\mathbf{c}), \quad \mathbf{c}\in\mathbb{C}^K,$$

where $\mathbf{B} = \mathbf{B}_1 + \mathbf{B}_2 + \mathbf{B}_3$ *with*

$$\begin{aligned}\mathbf{B}_1 &= \omega(\mathbb{E}\bar{x}_{k1}y_{k1}\bar{x}_{j1}y_{j1} - \rho_k\rho_j),\ 1\le k,j\le K,\\ \mathbf{B}_2 &= (\theta-\omega)(\mathbb{E}\bar{x}_{k1}y_{j1}\mathbb{E}\bar{x}_{j1}y_{k1}),\ 1\le k,j\le K,\\ \mathbf{B}_3 &= (\tau-\omega)(\mathbb{E}\bar{x}_{k1}\bar{x}_{j1}\mathbb{E}y_{j1}y_{k1}),\ 1\le k,j\le K,.\end{aligned}$$

Proof In the following, we only give the main steps of the proof and again interested readers are referred to Bai and Yao (2008) for the missing details.

First, as the base entries $\{x_{ij}\}$ and $\{y_{ij}\}$ have a finite 4-th moment we can first truncate these entries at $\delta_n\sqrt[4]{n}$, where $\delta_n\downarrow 0$, and then centralise them without affecting the limits.

Next, we use the moment convergence theorem to prove the lemma. Define

$$\xi_n = \sum_{l=1}^{K} c_l Z_l = \frac{1}{\sqrt{n}}\sum_{ij} a_{ij}\phi_{ij},$$

where

$$\phi_{ij} = \begin{cases}\sum_{l=1}^K c_l(\bar{x}_{il}y_{il} - \rho_l), & i = j,\\ \sum_{l=1}^K c_l\bar{x}_{il}y_{jl}, & i\ne j.\end{cases}$$

The goal is to find the limit of $\mathbb{E}\xi_n^h$. The power is expanded as in

$$\mathbb{E}\xi_n^h = n^{-h/2}\sum_{i_1,j_1,\dots,i_h,j_h=1}^{n} a_{i_1j_1}a_{i_2j_2}\cdots a_{i_h,j_h}\mathbb{E}\left(\phi_{i_1j_1}\phi_{i_2j_2}\cdots\phi_{i_h,j_h}\right).$$

Corresponding to each term in the sum, we draw a directed graph G consisting of h edges

$$i_1\to j_1,\ i_2\to j_2,\ \cdots,\ i_h\to j_h.$$

Define the notations

$$a_{i_1j_1}a_{i_2j_2}\cdots a_{i_h,j_h} = a_G, \qquad \phi_{i_1j_1}\phi_{i_2j_2}\cdots\phi_{i_h,j_h} = \phi_G.$$

Each graph G consists of several mutually non-connected subgraphs. These subgraphs fall into three categories:

- The first category consists of subgraphs which contain a single loop. Such a subgraph thus has only one vertex. Assume that we have m_1 subgraphs G_{1t} in this category and the degrees of the vertexes are $\mu_1, \dots, \mu_{m_1}$, respectively. Note that all μ_t must be positive even numbers. If there is some $\mu_t = 2$, the subgraph consists of one loop only and the corresponding expectation is 0. So, we can assume that $\mu_t \geq 4$ for all t. The contributions of the m_1 subgraphs are

$$\left| \prod_{t=1}^{m_1} \mathbb{E}\phi_{G_{1t}} \right| \leq K(\delta_n n^{1/4})^{\sum_{t=1}^{m_1}(\mu_t - 4)}.$$

- The second category of subgraphs contain at least one edge not belonging to a loop and one cycle. Assume that we have m_2 subgraphs in this category. Assume that the s-th subgraph G_{2s} contains u_s vertexes, the degrees are γ_{js}, $j = 1, \dots, u_s$ respectively. If there is some $\gamma_{js} = 1$, the value of the term will be 0. Therefore, the contribution of these m_2 connected subgraphs is bounded by

$$\left| \mathbb{E} \prod_{s=1}^{m_2} \phi_{G_{2s}} \right| \leq K(\delta_n \sqrt[4]{n})^{\sum_{s=1}^{m_2} \sum_{j=1}^{u_s}(\gamma_{js} - 2)}.$$

- The third category of subgraphs contain at least one edge not belonging to a loop but no cycle. Assume that we have m_3 subgraphs in this category. Assume that the s-th subgraph G_{3s} contains w_s vertexes, the degrees are ι_{js}, $j = 1, \dots, w_s$, respectively. Similarly, if there is some $\iota_{js} = 1$, the value of the term will be 0. Now, any vertex of the subgraph must have one loop, otherwise the value of the term equals to 0. The contribution of these m_3 subgraphs does not exceed

$$\left| \mathbb{E} \prod_{s=1}^{m_3} \phi_{G_{3s}} \right| \leq K(\delta_n \sqrt[4]{n})^{\sum_{s=1}^{m_3} \sum_{j=1}^{w_s}(\iota_{js} - 2) - 4I(m_3 \geq 0)}$$

Combining these three estimates, the total contribution from non-zero terms is bounded as

$$|\mathbb{E}\phi_G| \leq K(\delta_n \sqrt[4]{n})^{\sum_{t=1}^{m_1}(\mu_t - 4) + \sum_{s=1}^{m_2} \sum_{j=1}^{u_s}(\gamma_{js} - 2) + \sum_{s=1}^{m_3} \sum_{j=1}^{w_s}(\iota_{js} - 2) - 4I(m_3 \geq 0)}.$$

Next we estimate the coefficients a_G. For any w, $\sum_{j=1}^n |a_{jj}^w| \leq Kn$, so the contribution of the vertex in the first category of subgraphs does not exceed Kn^{m_1}. For a subgraph G_{2s} from the second category, assume that G_{2s} has t_s non-repeated edges $e_1, \dots, e_{t_s}$ and u_s different vertexes $v_1, \dots, v_{u_s}$, then we can choose one tree consisting of non-repeated edges, $e_1, \dots, e_{u_s - 1}$. Denote the tree by G_{2s1} and let the complement $G_{2s2} = G_{2s} - G_{2s1}$.

Note that when $e = (u, v)$, $u \neq v$,

$$\sum_v |a_{u,v}^2| \leq \|\mathbf{A}_n\|^2 \leq K,$$

We have

$$\sum_{v_1,\dots,v_{u_s}} |a_{G_{2s}}| = \sum_{v_1,\dots,v_{u_s}} \prod_{j=1}^{t_s} |a_{e_j}|$$
$$\leq \left(\sum_{v_1,\dots,v_{u_s}} \prod_{j=1}^{u_s - 1} |a_{e_j}|^2 \sum_{v_1,\dots,v_{u_s}} \prod_{j=u_s}^{t_s} |a_{e_j}^2| \right)^{1/2} \leq Kn^{(\nu_s + 1)/2},$$

where ν_s is the number of connected subgraph consisting of G_{2s2}. Obviously, $\nu_s \geq 1$. Here we use the fact: the contribution of first factor in the parenthesis is bounded by Kn. The contribution of the second factor does not exceed Kn^{ν_s}.

Similarly, for subgraphs in the third category,

$$\begin{aligned}\sum_{v_1,\ldots,v_{w_s}} |a_{G_{3s}}| &= \sum_{v_1,\ldots,v_{w_s}} \prod_{j=1}^{t_s} |a_{e_j}| \\ &\leq \left(\sum_{v_1,\ldots,v_{w_s}} \prod_{j=1}^{w_s-1} |a_{e_j}|^2 \sum_{v_1,\ldots,v_{w_s}} \prod_{j=w_s}^{t_s} |a_{e_j}^2| \right)^{1/2} \leq Kn^{w_s/2}.\end{aligned}$$

Finally, let $\mathcal{G}$ be the family of all the subgraphs whose contributions are non-negligible. Their sum can be bounded as follows:

$$\begin{aligned}& n^{-h/2} \sum_{G \in \mathcal{G}} a_G \mathbb{E}\,(\phi_G) \\ & \leq \sum^{*} Kn^{-\frac{h}{2}+m_1+\frac{1}{2}\sum_{2=1}^{m_2}(\nu_j+1)+\frac{1}{2}(\sum_{j=1}^{m_3} w_j+1)} \cdot (\delta_n \sqrt[4]{n})^{\sum(\mu_t-4)+\sum_{s=1}^{m_2}\sum_{j=1}^{u_s}(\gamma_{js}-2)+\sum_{s=1}^{m_3}\sum_{j=1}^{w_s}(\iota_{js}-2)-4I(m_3\geq 1)} \\ & = \sum^{*} Kn^{-\frac{1}{2}\sum_{s=1}^{m_2}(u_s-\nu_s-1)-\frac{1}{2}I(m_3\geq 1)} \cdot \delta_n^{\sum(\mu_t-4)+\sum_{s=1}^{m_2}\sum_{j=1}^{u_s}(\gamma_{js}-2)+\sum_{s=1}^{m_3}\sum_{j=1}^{w_s}(\iota_{js}-2)-4I(m_3\geq 1)}, \qquad (6.25)\end{aligned}$$

where $\sum^*$ is the sum of on the set defined by $\sum_{t=1}^{m_1} \mu_t + \sum_{s=1}^{m_2} \sum_{j=1}^{u_s} \gamma_{js} + \sum_{s=1}^{m_3} \sum_{j=1}^{w_s} \iota_{js} = 2h$. Obviously, for a term in (6.25) satisfying either $m_3 > 0$, or one of $\mu_t > 4$, or one of $\gamma_{js} > 2$, or one of $u_s > \nu_s + 1$, its contribution is negligible. So we need only to consider the situation of $m_3 = 0$, $\mu_t = 4$, $\gamma_{js} = 2$ and $u_s = \nu_s + 1$. Of course, when $\gamma_{js} = 2$, $\nu_s = 1$, which means $u_s = 2$. This implies $2m_1 + 2m_2 = h$. When h is odd number, this is impossible. So,

$$\mathbb{E}\xi_n^{2h+1} = o(1), \quad (h \geq 0).$$

For $\mathbb{E}\xi_n^{2h}$, we need only to consider the situation of $\mu_t = 4$, $u_s = 2$ and $\gamma_{js} = 2$. For each G_{1t}, it must be composed of $\mathbb{E}\phi_{jj}^2$. For each edge $e = (u, v)$ of G_{2s}, it is composed of $\mathbb{E}\phi_{uv}^2$ or $\mathbb{E}\phi_{uv}\phi_{vu}$. Assume that we have k_1 terms of type $\mathbb{E}\phi_{uv}^2$ and k_2 terms of type $\mathbb{E}\phi_{uv}\phi_{vu}$ in total. Then, we have

$$\begin{aligned}\mathbb{E}\xi_n^{2h} &= \sum_{m_1+k_1+k_2=h} \frac{(2h)!}{n^h 2^h m_1! k_1! k_2!} \left(\sum_{j=1}^{n} a_{jj}^2 \mathbb{E}\phi_{11}^2 \right)^{m_1} \\ & \quad \cdot \left(\sum_{u\neq v} a_{uv} a_{vu} \mathbb{E}(\phi_{12}\phi_{21}) \right)^{k_1} \left(\sum_{u \neq v} a_{uv}^2 \mathbb{E}(\phi_{12}^2) \right)^{k_2} + o(1) \\ &= \frac{(2h)!}{n^h 2^h h!} \left(\mathbb{E}\phi_{11}^2 \sum_{j=1}^{n} a_{jj}^2 + \mathbb{E}\phi_{12}\phi_{21} \sum_{u\neq v} |a_{uv}^2| + \mathbb{E}\phi_{12}^2 \sum_{u\neq v} a_{uv}^2 \right)^h + o(1). \qquad (6.26)\end{aligned}$$

Using elementary calculations leads to

$$\frac{1}{n} \left(\mathbb{E}\phi_{11}^2 \sum_{j=1}^{n} a_{jj}^2 + \mathbb{E}\phi_{12}\phi_{21} \sum_{u\neq v} |a_{uv}^2| + \mathbb{E}\phi_{12}^2 \sum_{u\neq v} a_{uv}^2 \right) = \frac{1}{2}(\mathbf{c}'\mathbf{B}\mathbf{c}) + o(1).$$

The conclusions of the lemma follow and details are skipped. □

While so far we have allowed various random variables be complex-valued, hereafter however we focus on the case of real variables in order to simplify the presentation.

A simple application of Lemma 6.7 yields the following CLT for random quadratic forms.

Theorem 6.8 *Let* $\{\mathbf{A}_n = [a_{ij}(n)]\}_n$ *be a sequence of* $n \times n$ *symmetric matrices satisfying the conditions of Lemma 6.7. Assume that* $\mathbf{w}_1, \cdots, \mathbf{w}_n$ *are iid* m*-dimensional real random vectors, with mean 0 and covariance matrix* $\mathbf{C} = (\sigma_{ij}) = \mathbb{E}[\mathbf{w}_1\mathbf{w}_1']$ *and a finite 4th-moment. Then, the random matrix*

$$\mathbf{R}_n = \frac{1}{\sqrt{n}}\left(W\mathbf{A}_nW' - \mathbf{C}\cdot tr\mathbf{A}_n\right), \quad where \quad \mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_n),$$

weakly converges to a $m \times m$*-dimensional symmetric random matrix* $\mathbf{R} = (R_{ij})$ *such that the joint distribution of* $\{\mathbf{R}_{i,j},\ i \leq j\}$ *is a* $\frac{1}{2}m(m+1)$*-dimensional Gaussian vector with mean 0 and covariance function*

$$cov(R_{ij}, R_{i'j'}) = \omega\left\{\mathbb{E}(w_{i1}w_{j1}w_{i'1}w_{j'1}) - \sigma_{ij}\sigma_{i'j'}\right\} + (\theta - \omega)\left\{\sigma_{ij'}\sigma_{i'j} + \sigma_{ii'}\sigma_{jj'}\right\}. \tag{6.27}$$

Proof Denote the m row vectors of $\mathbf{W}$ as $\mathbf{W}(i)$, $i = 1, \ldots, m$. The elements of the matrix $\mathbf{R}_n$ can also be seen as components of a random vector under an appropriate numbering

$$Z_n(l) = \frac{1}{\sqrt{n}}\left(\mathbf{X}_l'\mathbf{A}_n\mathbf{Y}_l - \rho_l\mathrm{tr}\mathbf{A}_n\right), \quad l = (i, j), \quad 1 \leq i \leq j \leq m,$$

such that when $Z_n(l) = \mathbf{R}_{n,ij}$, $\mathbf{X}_l = \mathbf{W}(i)$ and $\mathbf{Y}_l = \mathbf{W}(j)$. In particular, $\rho_l = \sigma_{ij}$. Application of Lemma 6.7 thus leads to the conclusions of the corollary. □

Corollary 6.9 *Assume that* $\mathbf{w}_1, \cdots, \mathbf{w}_n$ *are iid* m*-dimensional real random vectors, with mean 0 and covariance matrix* $\mathbf{C} = (\sigma_{ij}) = \mathbb{E}[\mathbf{w}_1\mathbf{w}_1']$ *and a finite 4th-moment. Then, the random matrix*

$$\mathbf{R}_n = \frac{1}{\sqrt{n}}\left(\sum_{i=1}^n \mathbf{w}_i\mathbf{w}_i' - n\mathbf{C}\right),$$

weakly converges to a $m \times m$*-dimensional symmetric random matrix* $\mathbf{R} = (R_{ij})$ *such that the joint distribution of* $\{\mathbf{R}_{i,j},\ i \leq j\}$ *is a* $\frac{1}{2}m(m+1)$*-dimensional Gaussian vector with mean 0 and covariance function*

$$cov(R_{ij}, R_{i'j'}) = \mathbb{E}(w_{i1}w_{j1}w_{i'1}w_{j'1}) - \sigma_{ij}\sigma_{i'j'}. \tag{6.28}$$

This result can be seen as a CLT for a (fixed) m-dimensional sample covariance matrix $\sum_i \mathbf{w}_i\mathbf{w}_i'$ with a population covariance matrix $\mathbf{C}$.

Theorem 6.10 *For the generalised spiked population model in* (6.5)*, assume that Conditions (i) to (vi) defined thereafter are valid and the base variables* $\{y_{ij}\}$ *are real. Then, the joint distribution of* L *random matrices defined in (6.22) converges to a multivariate Gaussian vector determined as follows: for any arbitrary* L *numbers* $a_1, \ldots, a_L$*, the random matrix*

$$\widetilde{\mathbf{R}}_n = a_1\mathbf{R}_n(l_1) + \cdots + a_L\mathbf{R}_n(l_L),$$

weakly converges to a Gaussian random matrix $\mathbf{R} = \{\mathbf{R}_{i,j},\ i \leq j\}$ *where*

(i) *the diagonal entries are i.i.d. zero-mean Gaussian with variance*

$$var(\mathbf{R}_{ii}) = \omega\left\{\mathbb{E}[|y_{i1}|^4] - 3\right\} + 2\theta; \tag{6.29}$$

(ii) *the upper off-diagonal entries are i.i.d. zero-mean Gaussian with variance θ; and*

(iii) *all these entries are mutually independent.*

Here the parameters θ and ω are

$$\theta = \sum_{j=1}^{L} a_j^2 \underline{s}'(l_j) + 2\sum_{j<k} a_j a_k \frac{\underline{s}(l_j) - \underline{s}(l_k)}{l_j - l_k}, \tag{6.30}$$

$$\omega = \left(\sum_{j=1}^{L} a_j \underline{s}(l_j)\right)^2. \tag{6.31}$$

Proof Recall the block matrices $\mathbf{Y}_1$ and $\mathbf{Y}_2$ defined in (6.8). We apply Theorem 6.8 by conditioning on $\mathbf{Y}_2$. Conditional on $\mathbf{Y}_2$, $\widetilde{\mathbf{R}}_n$ has the form of $\mathbf{R}_n$ in that theorem with $\mathbf{W} = \mathbf{Y}_1$, $\mathbf{C} = \mathbf{I}_m$ and

$$\mathbf{A}_n = a_1(l_1\mathbf{I}_n - \frac{1}{n}\mathbf{Y}_2^*\mathbf{V}_p\mathbf{Y}_2)^{-1} + \cdots + a_L(l_L\mathbf{I}_n - \frac{1}{n}\mathbf{Y}_2^*\mathbf{V}_p\mathbf{Y}_2)^{-1}.$$

To apply the theorem, we first verify the existence of the limits θ and ω for the sequence $\{\mathbf{A}_n\}$ as defined in Lemma 6.7, the limit τ being the same as θ for real variables. As l_j is outside the support of $F_{y,H}$, from Theorem 2.14,

$$\frac{1}{n}\mathrm{tr}(l_j\mathbf{I}_n - \frac{1}{n}\mathbf{Y}_2^*\mathbf{V}_p\mathbf{Y}_2)^{-1} \xrightarrow{a.s.} -\underline{s}(l_j),$$

$$\frac{1}{n}\mathrm{tr}(l_j\mathbf{I}_n - \frac{1}{n}\mathbf{Y}_2^*\mathbf{V}_p\mathbf{Y}_2)^{-2} \xrightarrow{a.s.} \underline{s}'(l_j),$$

$$\frac{1}{n}\mathrm{tr}(l_j\mathbf{I}_n - \frac{1}{n}\mathbf{Y}_2^*\mathbf{V}_p\mathbf{Y}_2)^{-1}(l_k\mathbf{I}_n - \frac{1}{n}\mathbf{Y}_2^*\mathbf{V}_p\mathbf{Y}_2)^{-1} \xrightarrow{a.s.} \frac{\underline{s}(l_j) - \underline{s}(l_k)}{l_j - l_k}.$$

Therefore, the limit in θ exists as

$$\frac{1}{n}\mathrm{tr}\mathbf{A}_n^2 \xrightarrow{a.s.} \theta = \sum_{j=1}^{L} a_j^2 \underline{s}'(l_j) + 2\sum_{j<k} a_j a_k \frac{\underline{s}(l_j) - \underline{s}(l_k)}{l_j - l_k}.$$

As for the limit ω, since the distribution of the distribution of the matrix $l_j\mathbf{I}_n - \frac{1}{n}\mathbf{Y}_2^*\mathbf{V}_p\mathbf{Y}_2$ is invariant under permutation of its rows, it is seen that the limit of its diagonal elements is the same as the limit of their average. Therefore,

$$\left(l_j\mathbf{I}_n - \frac{1}{n}\mathbf{Y}_2^*\mathbf{V}_p\mathbf{Y}_2\right)^{-1} \xrightarrow{\text{a.s.}} -\underline{s}(l_j),$$

and

$$\frac{1}{n}\sum_{j=1}^{n} a_{jj}^2 \xrightarrow{\text{a.s.}} \omega = \left(\sum_{j=1}^{L} a_j \underline{s}(l_j)\right)^2.$$

Next, for the covariance structure, since $\mathbf{C} = \mathbf{I}_m$, we have for the limiting matrix $\mathbf{R}$,

$$\mathrm{cov}(\mathbf{R}_{ij}, \mathbf{R}_{i'j'}) = \omega\left\{\mathbb{E}(y_{i1}y_{j1}y_{i'1}y_{j'1}) - \delta_{ij}\delta_{i'j'}\right\} + (\theta - \omega)\left\{\delta_{ij'}\delta_{i'j} + \delta_{ii'}\delta_{jj'}\right\}, \tag{6.32}$$

where $\delta_{\alpha\beta}$ is the Kronecker symbol. It is readily checked that

- For $i \neq i'$, $\text{cov}(\mathbf{R}_{ii}, \mathbf{R}_{i'i'}) = 0$ (between diagonal entries);
- $\text{var}(\mathbf{R}_{ii}) = \omega \left\{ \mathbb{E}[|y_{i1}|^4] - 3 \right\} + 2\theta$ (diagonal entries);
- For $i' < j'$, $\text{cov}(\mathbf{R}_{ii}, \mathbf{R}_{i'j'}) = 0$ (between diagonal and upper-diagonal entries);
- For $i < j$, $i' < j'$ and $(i, j) \neq (i', j')$, $\text{cov}(\mathbf{R}_{ij}, \mathbf{R}_{i'j'}) = 0$ (between upper-diagonal entries);
- For $i < j$, $\text{var}(\mathbf{R}_{ij}) = \theta$ (upper-diagonal entries).

The proof of Theorem 6.10 is complete. □

6.4.2 Derivation of the CLT for spiked sample eigenvalues

Let α_k be a fundamental spike eigenvalue. Following Theorem 6.3, the m_k packed sample eigenvalues $\{\lambda_j,\ j \in J_k\}$ are solutions of the equation $|\lambda - \mathbf{K}_n(\lambda)| = 0$ for large n and they converge almost surely to $\psi_k = \psi(\alpha_k)$. Also define the spectral decomposition of $\mathbf{\Lambda}$ to be

$$\mathbf{\Lambda} = \mathbf{U}\text{diag}(\alpha_1 \mathbf{I}_{m_1}, \ldots, \alpha_K \mathbf{I}_{m_K})\mathbf{U}^*, \tag{6.33}$$

where $\mathbf{U}$ is orthogonal.

By Eq. (6.21), we have

$$\lambda_j \mathbf{I} - \mathbf{K}_n(\lambda_j) = \lambda_j \mathbf{I} - \frac{\lambda_j}{\sqrt{n}} \mathbf{\Lambda}^{\frac{1}{2}} \mathbf{R}_n(\lambda_j) \mathbf{\Lambda}^{\frac{1}{2}} + \lambda_j \underline{s}(\lambda_j) \mathbf{\Lambda} + O_P(n^{-1}) \,.$$

Let

$$\delta_{n,j} = \sqrt{n}(\lambda_j - \psi_k) \,, \quad j \in J_k,$$

so that $\lambda_j = \psi_k + \delta_{n,j}/\sqrt{n}$. By Taylor expansion,

$$\underline{s}(\lambda_j) = \underline{s}(\psi_k) + \frac{\delta_{n,j}}{\sqrt{n}} \underline{s}'(\psi_k) + o_P(n^{-\frac{1}{2}}).$$

Therefore,

$$\mathbf{I} - \lambda_j^{-1} \mathbf{K}_n(\lambda_j) = \left\{ \mathbf{I} + \underline{s}(\psi_k) \mathbf{\Lambda} \right\} - \frac{1}{\sqrt{n}} \mathbf{\Lambda}^{\frac{1}{2}} \mathbf{R}_n(\lambda_j) \mathbf{\Lambda}^{\frac{1}{2}} + \frac{\delta_{n,j}}{\sqrt{n}} \underline{s}'(\psi_k) \mathbf{\Lambda} + o_P(n^{-\frac{1}{2}}). \tag{6.34}$$

By §6.4.1, the process $\{R_n(l),\ l \in \mathscr{U}$ weakly converges to a matrix-valued Gaussian process on $\mathscr{U}$. We now follow a method devised in Bai (1985) for limiting distributions of eigenvalues or eigenvectors from random matrices. First, we use Skorokhod strong representation so that on an appropriate probability space, the convergence of this process as well as (6.34) take place almost surely. Multiplying both sides of (6.34) by $\mathbf{U}^*$ from the left and by $\mathbf{U}$ from the right yields

$$\mathbf{U}[\mathbf{I} - \lambda_j^{-1} K_n(\lambda_j)]\mathbf{U}^* = \begin{pmatrix} \ddots & 0 & 0 \\ 0 & \{1 + \underline{s}(\psi_k)\alpha_u\} I_{m_u} & 0 \\ 0 & 0 & \ddots \end{pmatrix} - \frac{1}{\sqrt{n}} \mathbf{U}^* \mathbf{\Lambda}^{\frac{1}{2}} \mathbf{R}_n(\lambda_j) \mathbf{\Lambda}^{\frac{1}{2}} \mathbf{U}$$

$$+ \frac{1}{\sqrt{n}} \begin{pmatrix} \ddots & 0 & 0 \\ 0 & \delta_{n,j} \underline{s}'(\psi_k) \alpha_u I_{m_u} & 0 \\ 0 & 0 & \ddots \end{pmatrix} + o(n^{-\frac{1}{2}}) \,.$$

First, in the right side of the above, we see that all the non diagonal blocks tend to zero.

Next, for a diagonal block with index $u \neq k$, by definition $1 + \underline{s}(\psi_k)\alpha_u \neq 0$, and this is indeed the limit of that diagonal bloc since the contributions from the remaining three terms tend to zero. As for the the k-th diagonal block, $1 + \underline{s}(\psi_k)\alpha_k = 0$ by definition, the k-th diagonal block reduces to

$$-\frac{1}{\sqrt{n}}[\mathbf{U}^*\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{R}_n(\lambda_j)\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{U}]_{kk} + \frac{1}{\sqrt{n}}\delta_{n,j}\underline{s}'(\psi_k)\alpha_k I_{m_k} + o(n^{-\frac{1}{2}}).$$

For n sufficiently large, its determinant must be equal to zero,

$$\left|-\frac{1}{\sqrt{n}}[\mathbf{U}^*\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{R}_n(\lambda_j)\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{U}]_{kk} + \frac{1}{\sqrt{n}}\delta_{n,j}\underline{s}'(\psi_k)\alpha_k I_{m_k} + o(n^{-\frac{1}{2}})\right| = 0\ ,$$

or equivalently,

$$\left|-[\mathbf{U}^*\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{R}_n(\lambda_j)\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{U}]_{kk} + \delta_{n,j}\underline{s}'(\psi_k)\alpha_k I_{m_k} + o(1)\right| = 0.$$

Taking into account the convergence of the process $\{\mathbf{R}_n(l),\ l \in \mathscr{U}\}$ to $\{\mathbf{R}(l),\ l \in \mathscr{U}\}$ as defined in Theorem 6.10, it follows that $\delta_{n,j}$ tends to a solution of

$$\left|-[\mathbf{U}^*\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{R}(\psi_k)\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{U}]_{kk} + \lambda\underline{s}'(\psi_k)\alpha_k I_{m_k}\right| = 0.$$

that is, an eigenvalue of the $m_k \times m_k$ matrix

$$\frac{[\mathbf{U}^*\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{R}(\psi_k)\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{U}]_{kk}}{\alpha_k\underline{s}'(\psi_k)} = \frac{[\mathbf{U}^*\mathbf{R}(\psi_k)\mathbf{U}]_{kk}}{\underline{s}'(\psi_k)}\ .$$

Moreover, as the index j is arbitrary, all the J_k random variables $\sqrt{n}\{\lambda_j - \psi_k,\ j \leq J_k\}$ converge almost surely to the set of eigenvalues of the above matrix. Of cause, this convergence also holds in distribution on the new probability space, hence on the original one.

Finally for the calculation of the variances, let $\mathbf{G} = \mathbf{R}(\psi_k)/\underline{s}'(\psi_k)$. We have by Theorem 6.10,

$$\theta = \underline{s}'(\psi_k), \quad \omega = \underline{s}^2(\psi_k).$$

Let $\beta_y = (\mathbb{E}|y_{11}|^4 - 3)$. Then the variance of a diagonal element of $\mathbf{G}$ is

$$\sigma^2_{\alpha_k} = \frac{2\theta + \omega\beta_y}{\{\underline{s}'(\psi_k)\}^2} = \frac{2}{\underline{s}'(\psi_k)} + \beta\frac{\underline{s}^2(\psi_k)}{\{\underline{s}'(\psi_k)\}^2} = \alpha_k^2\psi'(\alpha_k)\{2 + \beta\psi'(\alpha_k)\}.$$

Here we have used the identities,

$$\underline{s}'(\psi(\alpha)) = \frac{1}{\alpha^2\psi'(\alpha)}\ , \quad \text{and} \quad \underline{s}(\psi(\alpha)) = -\frac{1}{\alpha}.$$

Summarising, we have proved the following theorem.

Theorem 6.11 *For the generalised spiked population model in* (6.5)*, assume that Conditions (i) to (iv) defined thereafter are valid and the random variables* $\{y_{ij}\}$*'s are real. Let* α_k *be a fundamental spike eigenvalue of multiplicity* m_k *and* $\{\lambda_i,\ i \in J_k\}$ *the corresponding spiked sample eigenvalues tending to* $\psi_k = \psi(\alpha_k)$*. Then the* m_k*-dimensional random vectors*

$$\sqrt{n}\{\ \lambda_i - \psi_k,\ i \in J_k\ \},$$

weakly converges to the distribution of eigenvalues of the $m_k \times m_k$ random matrix

$$\mathbf{M}_k = \mathbf{U}_k^* \mathbf{G}(\psi_k)\mathbf{U}_k, \tag{6.35}$$

where $\mathbf{U}_k$ is the k-th block of size $m_k \times m_k$ in $\mathbf{U}$ defined in (6.33) *corresponding to the spike α_k, and $G(\psi_k)$ is a Gaussian random matrix with independent entries such that*

(i) *its diagonal elements are i.i.d. Gaussian, with mean 0 and variance*

$$\sigma_{\alpha_k}^2 := \alpha_k^2 \psi'(\alpha_k)\{2 + \beta_y \psi'(\alpha_k)\}\ , \tag{6.36}$$

with $\beta_y = (\mathbb{E}|y_{11}|^4 - 3)$ denoting the fourth cumulant of the base entries y_{ij}'s;

(ii) *its upper triangular elements are i.i.d. Gaussian, with mean 0 and variance*

$$s_{\alpha_k}^2 := \alpha_k^2 \psi'(\alpha_k)\ . \tag{6.37}$$

In particular,

(i) *when the base entries $\{y_{ij}\}$ are Gaussian, $\beta_y = 0$, and then $\sigma_{\alpha_k}^2 = 2s_{\alpha_k}^2$: the matrix $\mathbf{G}$ is a real Gaussian Wigner matrix;*

(ii) *When the spike α_k is simple, i.e. $m_k = 1$, the limiting distribution of $\sqrt{n}\{\lambda_i - \psi_k\}$ is Gaussian.*

It is worth noticing that in case $\mathbf{\Lambda}$ is diagonal, $\mathbf{U} = \mathbf{I}_K$, the joint distribution of the k-th packed spiked sample eigenvalues is given by the eigenvalues of the Gaussian matrix G. This joint distribution is non-Gaussian unless the spike eigenvalue α_k is *simple*, i.e. $m_k = 1$. In this case with $\mathbf{U} = \mathbf{I}_K$,

$$\sqrt{n}\,(\lambda_i - \psi_k) \xrightarrow{\mathscr{D}} \mathcal{N}(0, \sigma_{\alpha_k}^2) \tag{6.38}$$

with the variance $\sigma_{\alpha_k}^2$ given in (6.36).

6.4.3 Examples and numeric illustrations of Theorem‘6.11

This section is devoted to describe in more details the content of Theorem 6.11 with several meaningful examples together with extended numerical computations. Throughout the section, we assume Johnstone's spiked population model so that the base LSD is the Marčenko-Pastur distribution F_y with index y and

$$\psi(\alpha) = \alpha + \frac{y\alpha}{\alpha - 1}, \quad \psi'(\alpha) = 1 - \frac{y}{(\alpha - 1)^2},$$

which is well defined for all spike eigenvalues $\alpha \neq 1$.

Example 6.12 *Gaussian variables where all spike eigenvalues are simple.* Assume that the variables $\{y_{ij}\}$ are real Gaussian, $\mathbf{\Lambda}$ is diagonal whose eigenvalues $\{\alpha_k\}$ are all simple. In other words, $K = m$ and $m_k = 1$ for all $1 \le k \le K$. Hence, $\mathbf{U} = I_m$. Following Theorem 6.11, for any spiked sample eigenvalue λ_i corresponding to a fundamental spike eigenvalue α_k,

$$\sqrt{n}(\lambda_{n,k} - \lambda_k) \xrightarrow{\mathscr{D}} \mathcal{N}(0, \sigma_k^2)$$

with the variance given in (6.36), i.e

$$\sigma_k^2 = 2\alpha_k^2 \psi'(\alpha_k) = \frac{2\alpha_k^2[(\alpha_k - 1)^2 - y]}{(\alpha_k - 1)^2}\ .$$

Example 6.13 *Gaussian variables with some multiple spike eigenvalues.* As in the previous example, the variables $\{y_{ij}\}$ are real Gaussian. Let $y = 0.5$ so that the Marčenko-Pastur distribution F_y has support $[a_y, b_y] = [0.086,\ 2.914]$. The critical interval for spike eigenvalues is $[1-\sqrt{y}, 1+\sqrt{y}] = [0.293,\ 1.707]$, *cf.* Figure 6.3.

Consider $K = 4$ spike eigenvalues $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (4, 3, 0.2, 0.1)$ with respective multiplicity $(m_1, m_2, m_3, m_4) = (1, 2, 2, 1)$. Let

$$\lambda_{n,1} \geq \lambda_{n,2} \geq \lambda_{n,3} \quad \text{and} \quad \lambda_{n,4} \geq \lambda_{n,5} \geq \lambda_{n,6}$$

be respectively, the three largest and the three smallest eigenvalues of the sample covariance matrix. Let as in the previous example

$$\sigma^2_{\alpha_k} = \frac{2\alpha_k^2[(\alpha_k - 1)^2 - y]}{(\alpha_k - 1)^2}\,. \tag{6.39}$$

We have $(\sigma^2_{\alpha_k}, k = 1, \ldots, 4) = (30.222,\ 15.75,\ 0.0175,\ 0.00765)$.

Following Theorem 6.11, we have

- For $j = 1$ and 6,
$$\delta_{n,j} = \sqrt{n}[\lambda_{n,j} - \phi(\alpha_k)] \xrightarrow{\mathscr{D}} \mathscr{N}(0, \sigma^2_{\alpha_k}). \tag{6.40}$$
Here, for $j = 1$, $k = 1$, $\phi(\alpha_1) = 4.667$ and $\sigma^2_{\alpha_1} = 30.222$; and for $j = 6$, $k = 4$, $\phi(\alpha_4) = 0.044$ and $\sigma^2_{\alpha_4} = 0.00765$.
- For $j = (2, 3)$ or $j = (4, 5)$, the two-dimensional vector $\delta_{n,j} = \sqrt{n}[\lambda_{n,j} - \phi(\alpha_k)]$ converges weakly to the distribution of (ordered) eigenvalues of the Gaussian Wigner random matrix
$$\mathbf{G} = \sigma_{\alpha_k}\begin{pmatrix} W_{11} & W_{12} \\ W_{12} & W_{22} \end{pmatrix},$$
where $\text{var}(W_{11}) = \text{var}(W_{22}) = 1$ and $\text{var}(W_{12}) = \frac{1}{2}$. Since the joint distribution of eigenvalues of a Gaussian Wigner matrix is known (see Mehta, 2004, e.g.), we get the following (non-ordered) density for the limiting distribution of $\delta_{n,j}$:
$$g(\delta, \gamma) = \frac{1}{4\sigma^3_{\alpha_k}\sqrt{\pi}}|\delta - \gamma| \exp\left[-\frac{1}{2\sigma^2_{\alpha_k}}(\delta^2 + \gamma^2)\right]. \tag{6.41}$$

Here is a numerical that compares the empirical distribution of the $\delta_{n,j}$'s to their limiting value. Dimensions are $p = 500$ and $n = 1000$. Empirical distributions of the six random variables $\{\delta_{n,j}, j = 1, \ldots, 6\}$ are obtained via 1000 independent simulations leading to

- a kernel density estimate for two univariate variables $\delta_{n,1}$ and $\delta_{n,6}$, denoted by $\widehat{f_{n,1}}$ $\widehat{f_{n,6}}$, respectively, and
- a kernel density estimate for two bivariate variables $(\delta_{n,2}, \delta_{n,3})$ and $(\delta_{n,4}, \delta_{n,5})$, denoted by $\widehat{f_{n,23}}$ $\widehat{f_{n,45}}$, respectively.

The kernel density estimates are computed using the R software implementing an automatic bandwidth selection method from Sheather and Jones (1991).

Figure 6.5 (left panel) compares the two univariate density estimates $\widehat{f_{n,1}}$ and $\widehat{f_{n,6}}$ with their Gaussian limits (6.40). As it can be seen, the empirical results confirm well the theoretic limit.

To compare the bivaiate density estimates $\widehat{f_{n,23}}$ and $\widehat{f_{n,45}}$ to their limiting densities given

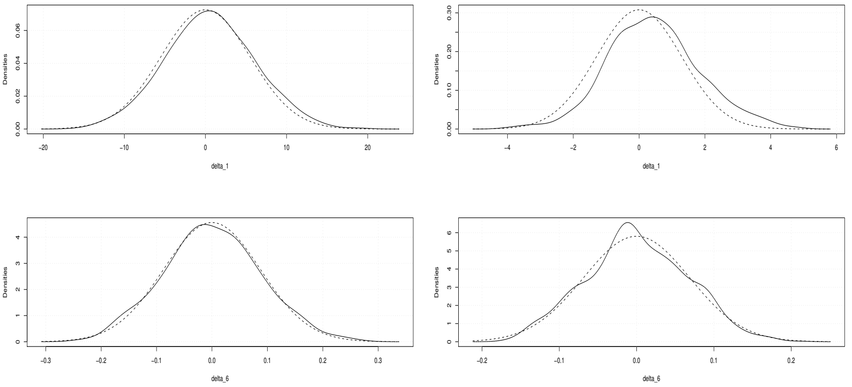


Figure 6.5 Empirical density estimates (in solid lines) from the largest (top: $\widehat{f_{n,1}}$) and the smallest (bottom: $\widehat{f_{n,6}}$) sample eigenvalue from 1000 independent replications, compared to their Gaussian limits (dashed lines). Dimensions are $p = 500$ and $n = 1000$. Left panel: Gaussian entries. Right panel: binary entries.

in (6.41), their contour lines are displayed in Figure 6.6, left panel, for $\widehat{f_{n,23}}$ and Figure 6.7, left panel, for $\widehat{f_{n,45}}$. Again the theoretical result is well confirmed.

Example 6.14 *A case with binary entries.*

As in the previous example, this example uses $y = 0.5$ and the same spike eigenvalues $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (4, 3, 0.2, 0.1)$ with multiplicities $(m_1, m_2, m_3, m_4) = (1, 2, 2, 1)$. Assume again that $\mathbf{\Lambda}$ is diagonal but this time binary entries $\{y_{ij}\}$ are considered. That is, both $\mathbf{y}_{1i}$ and and $\mathbf{y}_{2i}$ are made with i.i.d. binary variables $\{\varepsilon_j\}$ taking values in $\{+1, -1\}$ with equiprobability. Notice that $\mathbb{E}\varepsilon_j = 0$, $\mathbb{E}\varepsilon_j^2 = 1$ and $\beta_y = -2$. This non-null value denotes a departure from the Gaussian case.

As in the previous example, the limiting distributions of the three largest and the three smallest eigenvalues $\{\lambda_{n,j},\ j = 1, \ldots, 6\}$ of the sample covariance matrix are examined. Following Theorem 6.11,

- For $j = 1$ and 6,

$$\delta_{n,j} = \sqrt{n}[\lambda_{n,j} - \phi(\alpha_k)] \xrightarrow{\mathscr{D}} \mathscr{N}(0, s_{\alpha_k}^2), \quad s_{\alpha_k}^2 = \sigma_{\alpha_k}^2 \frac{y}{(\alpha_k - 1)^2},$$

  where $\sigma_{\alpha_k}^2$ is the limiting variance in (6.39) for the case with Gaussian entries. Compared to the previous Gaussian case, the additional factor $y/(\alpha_k - 1)^2 < 1$ (since a fundamental spike is such that $|\alpha_k - 1| > \sqrt{y}$), so that the limiting Gaussian distributions of the largest and the smallest eigenvalue are less dispersed.
- For $j = (2, 3)$ or $j = (4, 5)$, the two-dimensional vector $\delta_{n,j} = \sqrt{n}[\lambda_{n,j} - \phi(\alpha_k)]$ converges weakly to the distribution of (ordered) eigenvalues of the Gaussian random matrix

$$\mathbf{G} = \sigma_{\alpha_k} \begin{pmatrix} W_{11} & W_{12} \\ W_{12} & W_{22} \end{pmatrix}. \tag{6.42}$$

Here, $\mathrm{var}(W_{12}) = \frac{1}{2}$ as previously but $\mathrm{var}(W_{11}) = \mathrm{var}(W_{22}) = y/(\alpha_k - 1)^2 < 1$. Therefore, the matrix $W = (W_{ij})$ is no more a real Gaussian Wigner matrix. Unlike the

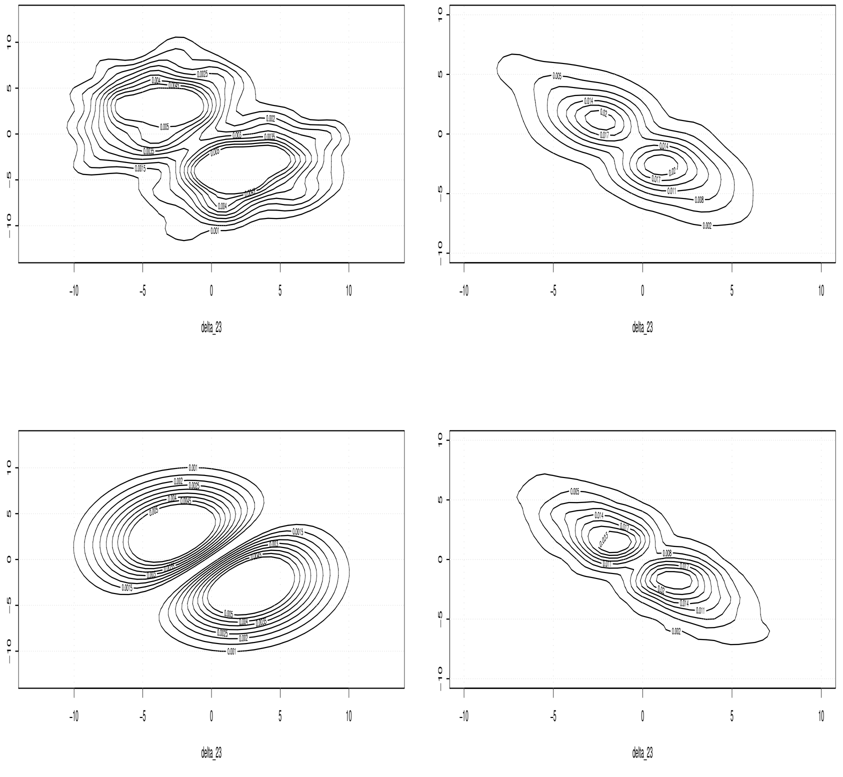


Figure 6.6 Limiting bivariate distribution from the second and the third sample eigenvalues. Left panel: Gaussian entries; right panel: binary entries. Top: contour lines of the empirical kernel density estimates $\widehat{f_{n,23}}$ from 1000 independent replications with $p = 500$, $n = 1000$. Bottom: Contour lines of their limiting distribution, given by the eigenvalues of a 2×2 Gaussian Wigner matrix on the left panel and and computed using 10000 independent replications on the right panel.

previous Gaussian case, the joint distribution of the eigenvalues of W is unknown analytically. Here it is computed empirically by simulating this joint density using 10000 independent replications. Again, as $y/(\alpha_k - 1)^2 < 1$, these limiting distributions are less dispersed than in the previous Gaussian case.

The kernel density estimates $\widehat{f_{n,1}}$, $\widehat{f_{n,6}}$, $\widehat{f_{n,23}}$ and $\widehat{f_{n,45}}$ are computed as in the previous case using $p = 500$, $n = 1000$ and 1000 independent replications.

Figure 6.5 (right panel) compares the two univariate density estimates $\widehat{f_{n,1}}$ and $\widehat{f_{n,6}}$ to their Gaussian limits. Again, empirical results confirm well the theoretical limit. Notice however a seemingly slower convergence to the Gaussian limit in this case than in the previous case with Gaussian entries.

The bivaiate density estimates $\widehat{f_{n,23}}$ and $\widehat{f_{n,45}}$ are then compared to their limiting densities as previously in the right panels of Figures 6.6 and 6.7, respectively. The shapes of these bivariate limiting distributions are rather different from the previous Gaussian

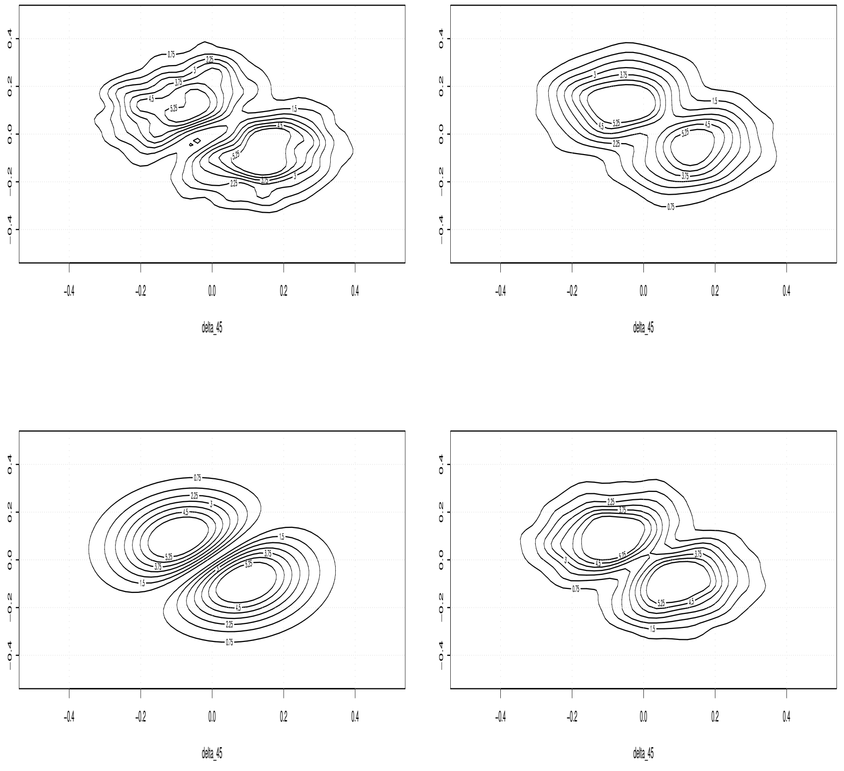


Figure 6.7 Limiting bivariate distribution from the second and the third smallest sample eigenvalues. Limiting bivariate distribution from the second and the third sample eigenvalues. Left panel: Gaussian entries; right panel: binary entries. Top: contour lines of the empirical kernel density estimates $\widehat{f_{n,45}}$ from 1000 independent replications with $p = 500$, $n = 1000$. Bottom: Contour lines of their limiting distribution, given by the eigenvalues of a 2×2 Gaussian Wigner matrix on the left panel and computed using 10000 independent replications on the right panel.

case. Notice that the limiting bivariate densities are obtained by simulations of 10000 independent G matrices given in (6.42) in this case with binary entries.

6.5 Estimation of the values of spike eigenvalues

Previous discussions in the chapter show that for a fundamental spike eigenvalue α_k of multiplicity m_k, there are m_k packed sample eigenvalues $\{\lambda_i, i \in J_k\}$ that converge almost surely to $\psi_k = \psi(\alpha_k)$. Moreover, a related CLT is given in Theorem 6.11. For statistical applications, it is of primary importance to estimate the spike eigenvalues $\{\alpha_k\}$. For example in principal component analysis, the contribution of the k-th principal component is given by $m_k\alpha_k/\operatorname{tr}(\mathbf{\Sigma})$.

6.5.1 Estimation when the ψ-function is known

If the functional form of ψ is known, a natural estimator of λ_k is just $\hat{\alpha}_k = \psi^{-1}(\lambda_i)$. For example, for Johnstone's spiked population model, $\psi(\alpha) = \alpha + y\alpha/(\alpha - 1)$ is known and this inversion method provides a way for the estimation of the spike eigenvalues $\{\alpha_k\}$.

Theorem 6.15 *Assume that the function ψ is known, define*

$$\hat{\alpha}_k = \frac{1}{m_k} \sum_{i \in J_k} \psi^{-1}(\lambda_i) . \tag{6.43}$$

Assume the same conditions as in Theorem 6.11 hold. Then

$$\sqrt{n}(\hat{\alpha}_k - \alpha_k) \xrightarrow{\mathscr{D}} \frac{1}{m_k \psi'(\psi_k)} tr\left[\mathbf{U}_k^* \mathbf{G}(\psi_k) \mathbf{U}_k\right], \tag{6.44}$$

which is a centred Gaussian distribution and where $\mathbf{U}_k$ and $\mathbf{G}(\psi_k)$ are defined in (6.35).

Proof From Taylor expansion and by Theorem 6.11,

$$\begin{aligned}\sqrt{n}(\hat{\alpha}_k - \alpha_k) &= \frac{\sqrt{n}}{m_k} \sum_{i \in J_k} (\psi^{-1}(\lambda_i) - \psi^{-1}(\psi_k)) \\ &= \frac{\sqrt{n}}{m_k \psi'(\alpha_k)} \sum_{i \in J_k} (\lambda_i - \psi_k) + o_p(1) \\ &\xrightarrow{\mathscr{D}} \frac{1}{m_k \psi'(\alpha_k)} \mathrm{tr}\left[\mathbf{U}_k^* \mathbf{G}(\psi_k) \mathbf{U}_k\right],\end{aligned}$$

where $\mathbf{U}_k$ and the Gaussian limiting matrix $\mathbf{G}(\psi_k)$ are given in (6.35). Clearly, this distribution is a centred Gaussian. □

6.5.2 Estimation when the ψ-function is unknown

In most of practical cases however, the function ψ is unknown or is much too complicate for the purpose of estimation of the spike eigenvalues. In this section, a new estimator is proposed.

The method is based on the fundamental equation (6.14) which states that when a spiked sample eigenvalue λ_j converges to $\psi_k = \psi(\alpha_k)$ for some fundamental spike λ_k, then by definition it holds that

$$\alpha_k \underline{s}(\psi_k) = -1 .$$

As the companion Stieltjes transform $\underline{s}$ of the LSD $F_{y,H}$ can be consistently estimated from sample eigenvalues, plugging such an estimate in the above equation leads to a consistent estimate of α_k.

More precisely, for each $z \notin \Gamma_{F_{y,H}}$, the Stieltjes transform of the ESD of a sample covariance matrix $\mathbf{S}_n$ *without spike eigenvalues*

$$s_n(z) = \frac{1}{p} \sum_{i=1}^{p} \frac{1}{\lambda_i - z}$$

converge almost surely to the Stieltjes transform s of $F_{y,H}$. Here $(\lambda_i)_{i=1}^p$ denotes the p eigenvalues of $\mathbf{S}_n$. When the population has (a finite number of) spike eigenvalues, the above still hold if we exclude in the sum the spiked sample eigenvalues, that is with $J = \cup_{k=1}^{K} J_k$,

$$s_n^*(z) = \frac{1}{p} \sum_{i \notin J} \frac{1}{\lambda_i - z} \xrightarrow{\text{a.s.}} s(z), \quad z \notin \Gamma_{F_{y,H}} .$$

The same holds true for the companion Stieltjes transforms, denoting $y_n = p/n$,

$$\underline{s}_n^*(z) = -\frac{1 - y_n}{z} + \frac{1}{n} \sum_{i \notin J} \frac{1}{\lambda_i - z} \xrightarrow{\text{a.s.}} \underline{s}(z). \tag{6.45}$$

Therefore, for λ_j converging to ψ_k where $j \in J_k$, define

$$m_{n,j} = -\frac{1 - y_n}{\lambda_j} + \frac{1}{n} \sum_{i \notin J} \frac{1}{\lambda_i - \lambda_j} . \tag{6.46}$$

So $m_{n,j}$ is an analogue of $\underline{s}_n^*$ evaluated at the spiked sample eigenvalue λ_j.

The theorem below establishes the consistency of $-1/m_{n,j}$ as an estimator of α_k.

Theorem 6.16 *Let α_k be a fundamental spike eigenvalue from the generalised spiked population model as defined in §6.2 and satisfying Conditions (i)-(vi). For any $j \in J_k$ and the associated spiked sample eigenvalues λ_j, define $m_{n,j}$ as in Eq.* (6.46)*. Then,*

$$-\frac{1}{m_{n,j}} \xrightarrow{a.s.} \alpha_k. \tag{6.47}$$

Proof Since $\psi_k \notin \Gamma_{F_{y,H}}$, we know already that $\underline{s}_n^*(\psi_k) \xrightarrow{\text{a.s.}} \underline{s}(\psi_k)$ and then $-1/\underline{s}_n^*(\psi_k) \xrightarrow{\text{a.s.}} -1/\underline{s}(\psi_k) = \alpha_k$. The difference between $m_{n,j}$ and $\underline{s}_n^*(\psi_k)$ is

$$m_{n,j} - \underline{s}_n^*(\psi_k) = (1 - y_n)\left(\frac{1}{\psi_k} - \frac{1}{\lambda_j}\right) + (\lambda_j - \psi_k)\frac{1}{n} \sum_{i \notin J} \frac{1}{(\lambda_i - \lambda_j)(\lambda_i - \psi_k)}.$$

We have almost surely,

$$\liminf \left\{\inf_{i \notin J} |\lambda_i - \lambda_j|\right\} \geq \delta > 0, \quad \liminf \left\{\inf_{i \notin J} |\lambda_i - \psi_k|\right\} \geq \delta > 0,$$

for some positive constant δ. Therefore, since $\lambda_j \xrightarrow{\text{a.s.}} \psi_k$, $m_{n,j} - \underline{s}_n^*(\psi_k) \xrightarrow{\text{a.s.}} 0$ and $-1/m_{n,j} \xrightarrow{\text{a.s.}} -1/\underline{s}(\psi_k) = \alpha_k$. □

6.6 Estimation of the number of spike eigenvalues

In §6.5, estimators are proposed for fundamental spike eigenvalues. These estimators rely on the fact that the separation between spike eigenvalues and base eigenvalues is completely known, that is one knows in advance that there are K fundamental eigenvalues $\{\alpha_k\}$ with respective multiplicity number m_k $(1 \leq k \leq K)$. In real-life data analysis as in the two examples given at the beginning of the chapter, such information is not available and it has to be inferred from the data either.

The spiked population model is naturally connected to the following *signal detection* model. Signals are recorded using p recorders in order to detect an unknown number of *m source signals*. As a first approximation, the recorded signals can be thought as linear combinations of the *source signals*. If we denote by $\mathbf{x}_t = (x_{t1}, \dots, x_{tp})'$ the p signals recorded at time t, and by $\mathbf{s}_t = (x_{t1}, \dots, x_{tp})'$ the source signals emitted at time t, we have

$$\mathbf{x}_t = \mathbf{A}\mathbf{s}_t + \boldsymbol{\varepsilon}_t \ , \tag{6.48}$$

where $\mathbf{A}$ is a $p \times m$ *mixing matrix* representing the source-recording mechanism and $\boldsymbol{\varepsilon}_t$ a measurement error. It is reasonable to assume that (i) the noise and the source signal are independent; (ii) the noise is centred with a covariance matrix $\operatorname{cov}(\boldsymbol{\varepsilon}_t) = \sigma^2 \mathbf{I}_p$. Then

$$\boldsymbol{\Sigma} = \operatorname{cov}(\mathbf{x}_t) = \mathbf{A} \operatorname{cov}(\mathbf{s}_t)\mathbf{A}' + \sigma^2 \mathbf{I}_p \ .$$

It is clear that the rank of $\mathbf{A} \operatorname{cov}(\mathbf{s}_t)\mathbf{A}'$ does not exceed m and if we denote its eigenvalues by α_j with respective multiplicity numbers m_j ($\sum m_j = m$), then clearly

$$\operatorname{spec}(\boldsymbol{\Sigma}) = (\underbrace{\alpha_1, \dots, \alpha_1}_{m_1}, \dots, \underbrace{\alpha_K, \dots, \alpha_K}_{m_K}, \underbrace{0, \dots, 0}_{p-m}) + \sigma^2(\underbrace{1, \dots, 1}_{p}). \tag{6.49}$$

If we rewrite the spectrum as

$$\operatorname{spec}(\boldsymbol{\Sigma}) = \sigma^2(\underbrace{\alpha'_1, \dots, \alpha'_1}_{m_1}, \dots, \underbrace{\alpha'_K, \dots, \alpha'_K}_{m_K}, \underbrace{1, \cdots, 1}_{p-m}), \tag{6.50}$$

it is readily seen that the model coincide with Johnstone's spiked population model introduced in §6.2.1 where simply, all the variances are multiplied by the noise variance σ^2. Finding the number m of spikes, or *detecting the number m of signals*, constitutes an unavoidable step before envisaging other inference tasks such as the estimation of the mixing matrix $\mathbf{A}$ or estimation of the signal strengths α_i ($i = 1, \dots, K$).

6.6.1 The estimator

Assume for the moment that all the spike eigenvalues $(\alpha_k + \sigma^2)_{1 \le i \le m}$ are simple, i.e. $m_k = 1$ and $K = m$. Moreover, assumed that $\alpha'_1 > \cdots > \alpha'_m > 1 + \sqrt{y}$ for all $i \in \{1, \dots, m\}$; i.e all the spike eigenvalues are fundamental so that the corresponding spiked sample eigenvalues converge to a limit outside the support of the LSD. Again let $\lambda_1 \ge \cdots \ge \lambda_p$ be the sample eigenvalues from $\mathbf{S}_n$. By Corollary 6.4, for each $1 \le k \le m$,

$$\lambda_k \xrightarrow{\text{a.s.}} \sigma^2 \psi(\alpha'_k),$$

and for all $1 \le i \le L$ with a prefixed range L,

$$\lambda_{m+i} \xrightarrow{\text{a.s.}} b = \sigma^2(1 + \sqrt{y})^2.$$

It is thus possible to built an estimator for m following a close inspection of differences between consecutive eigenvalues

$$\delta_j = \lambda_j - \lambda_{j+1}, \ j \ge 1.$$

Indeed, the results quoted above imply that a.s. $\delta_j \to 0$, for $j \ge m$ whereas for $j < m$, δ_j tends to a positive limit. Thus it becomes possible to estimate m from index-numbers j where δ_j become small. More precisely, define the estimator of m to be

$$\hat{q}_n = \min\{j \in \{1, \dots, s\} : \delta_{j+1} < d_n\}, \tag{6.51}$$

where $s > m$ is a fixed number big enough, and d_n is a threshold to be defined. In practice, s should be thought as a preliminary bound on the number of possible spikes. In fact, in this case where all the spikes are simple, it can be proved that $\hat{q}_n \xrightarrow{\mathscr{P}} m$ providing that the threshold satisfies $d_n \to 0$, $n^{2/3}d_n \to +\infty$ and under mild assumptions on the moments of the base variables $\{y_{ij}\}$.

When some of spikes, say α_k, are multiple, then $\delta_j = \lambda_j - \lambda_{j+1} \xrightarrow{\text{a.s.}} 0$ when spiked sample eigenvalues λ_j and λ_{j+1} are both associated to α_k, i.e. $\{j, j+1\} \subset J_k$. This fact creates an ambiguity with those differences δ_j corresponding to the noise eigenvalues, i.e. $j \geq m$, which also tend to zero. However, the convergence of the δ_j's, for $j > m$ (noise) is faster (in $O_{\mathbb{P}}(n^{-2/3})$) than that of the δ_j from a same spike eigenvalue α_k (in $O_{\mathbb{P}}(n^{-1/2})$) following the CLT in Theorem 6.11. This is the key feature that allows a suitable choice of the threshold d_n to guarantee the consistence of the estimator $\hat{q}_n$.

Theorem 6.17 *Consider Johnstone's spiked population model satisfying the assumptions (i)-(vi) introduced in §6.2 where the covariance matrix has the structure given in* (6.50). *Moreover, the variables $\{y_{ij}\}$ are assumed to have a symmetric law and a sub-exponential decay, that means there exists positive constants D, D′ such that, for all $t \geq D'$,*

$$\mathbb{P}(|y_{ij}| \geq t^D) \leq e^{-t}.$$

Let $(d_n)_{n\geq 0}$ be a real sequence such that $d_n = o(n^{-1/2})$ and $n^{2/3}d_n \to +\infty$. Then the estimator $\hat{q}_n$ is consistent, i.e $\hat{q}_n \to m$ in probability when $n \to +\infty$.

Proof Without loss of generality we can assume that $\sigma^2 = 1$ (if it is not the case, we consider λ_j/σ^2). The proof is based on the following key properties:

(i) By Theorem 6.11, for a fundamental spike eigenvalue α_k, the m_k packed eigenvalues

$$\sqrt{n}[\lambda_j - \psi(\alpha_k')],\ j \in J_k\ , \tag{6.52}$$

has a limiting distribution;

(ii) a technical result stating that for all $1 \leq i \leq L$ with a prefixed range L,

$$n^{\frac{2}{3}}(\lambda_{m+i} - b) = O_{\mathbb{P}}(1), \tag{6.53}$$

which is a direct consequence of Proposition 5.8 of Benaych-Georges et al. (2011).

We have

$$\begin{aligned}\{\hat{q}_n = m\} &= \{m = \min\{j : \delta_{j+1} < d_n\}\} \\ &= \{\forall j \in \{1, \dots, m\}, \delta_j \geq d_n\} \cap \{\delta_{m+1} < d_n\}.\end{aligned}$$

Therefore

$$\begin{aligned}\mathbb{P}(\hat{q}_n = m) &= \mathbb{P}\left(\bigcap_{1\leq j\leq m} \{\delta_j \geq d_n\} \cap \{\delta_{m+1} < d_n\}\right) \\ &= 1 - \mathbb{P}\left(\bigcup_{1\leq j\leq m} \{\delta_j < d_n\} \cup \{\delta_{m+1} \geq d_n\}\right) \\ &\geq 1 - \sum_{j=1}^{m} \mathbb{P}(\delta_j < d_n) - \mathbb{P}(\delta_{m+1} \geq d_n).\end{aligned}$$

Case of $j = m + 1$. In this case, $\delta_{m+1} = \lambda_{m+1} - \lambda_{m+2}$ (noise eigenvalues). As $d_n \to 0$ such that, $n^{2/3} d_n \to +\infty$, and by using (6.53),

$$\mathbb{P}(\delta_{m+1} \geq d_n) \to 0.$$

Case of $1 \leq j \leq m$. These indexes correspond to the spike eigenvalues.

- Let $I_1 = \{1 \leq l \leq m | \text{card}(J_l) = 1\}$ (simple spike) and $I_2 = \{l - 1 | l \in I_1 \text{ and } l - 1 > 1\}$. For all $j \in I_1 \cup I_2$, δ_j corresponds to a consecutive difference of λ_j issued from two different spikes, it follows that, again using (6.52), it can be proved that

$$\mathbb{P}(\delta_j < d_n) \to 0, \forall j \in I_1.$$

- Let $I_3 = \{1 \leq l \leq m - 1 | l \notin (I_1 \cup I_2)\}$. For all $j \in I_3$, it exists $k \in \{1, \ldots, K\}$ such that $j \in J_k$.
  - If $j + 1 \in J_k$ then, by (6.52), $\mathbf{X}_n = \sqrt{n}\delta_j$ converges weakly to a limit which has a density function on $\mathbb{R}^+$. So by using Lemma 6.18 below and that $d_n = o(n^{-1/2})$, we have

$$\mathbb{P}\left(\delta_j < d_n\right) = \mathbb{P}\left(\sqrt{n}\delta_j < \sqrt{n} d_n\right) \to 0;$$

  - Otherwise, $j + 1 \notin J_k$, so $\alpha_j \neq \alpha_{j+1}$. Consequently, as previously, δ_j corresponds to a consecutive difference of λ_j issued from two different spikes, and it can be proved as previously using (6.52), that

$$\mathbb{P}(\delta_j < d_n) \to 0.$$

- The case of $j = m$ can be treated in a similar way, thus omitted.

In conclusion, $\mathbb{P}(\delta_{m+1} \geq d_n) \to 0$ and $\sum_{j=1}^{m} \mathbb{P}(\delta_j < d_n) \to 0$; it then follows that $\mathbb{P}(\hat{q}_n = m) \to 1$. □

Lemma 6.18 *Let* $(\mathbf{X}_n)_{n\geq 0}$ *be a sequence of positive random variables which weakly converges to a probability distribution with a continuous cumulative distribution function. Then for all real sequence* $(u_n)_{n\geq 0}$ *which converges to 0,*

$$\mathbb{P}(\mathbf{X}_n \leq u_n) \to 0.$$

Proof As $(\mathbf{X}_n)_{n\geq 0}$ converges weakly, there exists a function G such that, for all $v \geq 0$, $\mathbb{P}(\mathbf{X}_n \leq v) \to G(v)$. Furthermore, as $u_n \to 0$, there exists $N \in \mathbb{N}$ such that for all $n \geq N$, $u_n \leq v$. So $\mathbb{P}(\mathbf{X}_n \leq u_n) \leq \mathbb{P}(\mathbf{X}_n \leq v)$, and $\overline{\lim}_{n\to+\infty} \mathbb{P}(\mathbf{X}_n \leq u_n) \leq \overline{\lim}_{n\to+\infty} \mathbb{P}(\mathbf{X}_n \leq v) = G(v)$. Now we can take $v \to 0$: as $(\mathbf{X}_n)_{n\geq 0}$ is positive, $G(v) \to 0$. Consequently, $\mathbb{P}(\mathbf{X}_n \leq u_n) \to 0$. □

There is a variation of the estimator defined as follows. Instead of making a decision once one difference δ_k is below the threshold d_n, see (6.51), one may decide once two consecutive differences δ_k and δ_{k+1} are both below d_n, i.e. define the estimator to be

$$\hat{q}_n^* = \min\{j \in \{1, \ldots, s\} : \delta_{j+1} < d_n \text{ and } \delta_{j+2} < d_n\}. \tag{6.54}$$

It can be easily checked that the proof for the consistency of $\hat{q}_n$ applies equally to $\hat{q}_n^*$ under the same conditions as in Theorem 6.17. This version of the estimator will be used in all the simulation experiments below. Intuitively, $\hat{q}_n^*$ should be more robust than $\hat{q}_n$. Notice that eventually more than two consecutive differences could be used in (6.54). However, simulation experiments reported below show that using more consecutive differences does not improve significantly.

6.6.2 Implementation issues and overview of simulation experiments

The practical implementation of the estimator $\hat{q}_n^*$ depend on two unknown parameters, namely the noise variance σ^2 and the threshold sequence d_n. Simulation experiments use an improved version of the following maximum likelihood estimator

$$\widehat{\sigma^2} = \frac{1}{p-m} \sum_{i=m+1}^{p} \lambda_i. \tag{6.55}$$

It remains to choose a threshold sequence d_n. The choice here is of the form $d_n = Cn^{-2/3}\sqrt{2\log\log n}$, where C is a "tuning" parameter to be adjusted. In the Monte-Carlo experiments below, two choices of C are considered: the first one is manually tuned and used to assess some theoretical properties of the estimator $\hat{q}_n^*$; and the second one is a data-driven and automatically calibrated one which is detailed in §6.6.3.

In the remaining of the section, extensive simulation experiments are conducted to assess the quality of the estimator $\hat{q}_n^*$ including a detailed comparison with a benchmark detector known as Kritchman and Nadler's detector (KN).

In all experiments, data are generated with the assigned noise level $\sigma^2 = 1$ and empirical values are calculated using 500 independent replications. Table 6.1 gives a summary of the design in the experiments. One should notice that both the given value of $\sigma^2 = 1$ and the estimated one, as well as the manually tuned and the automatic chosen values of C are used in different scenarios. There are in total three sets of experiments. The first set (Figures 6.8 and 6.9 and Models A, B), given in this section, illustrates the convergence of the estimator $\hat{q}_n^*$. The second set of experiments (Figures 6.10 and 6.11 and Models D-K) addresses the performance of the automatic tuned C and they are reported in §6.6.3. The last set of experiments (Figures 6.12, 6.13 and 6.14), reported in §6.6.4, are designed for a comparison with the benchmark detector KN.

Multiple spikes versus simple spikes

In Figure 6.8, the case of a single spike α is considered and the probability of misestimation analysed as a function of the value of α, for $(p, n) = (200, 800)$, $y = 0.25$ and $(p, n) = (2000, 500)$, $y = 4$. For the first case $C = 5.5$ and for the second case $C = 9$ (all manually tuned). The noise level $\sigma^2 = 1$ is given. The estimator $\hat{q}_n^*$ performs well; in particular, the critical level $\sqrt{y}$ from which the behaviour of the spike eigenvalues differs from the noise ones is recovered ($\sqrt{y} = 0.5$ for the first case, and 2 for the second).

Next, the same parameters are used with addition of some multiple spikes. Figure 6.9 concerns Model A: $(\alpha_1, \alpha_2, \alpha_3) = (\alpha, \alpha, 5)$, $0 \leq \alpha \leq 2.5$ and Model B: $(\alpha_1, \alpha_2, \alpha_3) = (\alpha, \alpha, 15)$, $0 \leq \alpha \leq 8$. The dimensions are $(p, n) = (200, 800)$ and $C = 6$ for Model A, and $(p, n) = (2000, 500)$ and $C = 9.9$ for Model B.

There is no particular difference with the previous situation: when spikes are close or even equal, or near to the critical value, the estimator remains consistent although the convergence rate becomes slower. Overall, the estimator $\hat{q}_n^*$ is able to find the number of spikes.

Table 6.1 *Summary of parameters used in the simulation experiments. (L: left, R: right)*

Fig.	Mod.	spike		Fixed parameters			Var.
No.	No.	values	p, n	y	σ^2	C	par.
1		(α)	$(200, 800)$ $(2000, 500)$	1/4 4	Given	5.5 9	α
2	A B	$(\alpha, \alpha, 5)$ $(\alpha, \alpha, 15)$	$(200, 800)$ $(2000, 500)$	1/4 4	Given	6 9.9	α
3L	D E	$(6, 5)$ $(6, 5, 5)$		10	Given	11 and auto	n
3R	F G	$(10, 5)$ $(10, 5, 5)$		1	Given	5 and auto	n
4L	H I	(1.5) $(1.5, 1.5)$		1	Given	5 and auto	n
4R	J K	$(2.5, 1.5)$ $(2.5, 1.5, 1.5)$		1	Given	5 and auto	n
5L	D	$(6, 5)$		10	Estimated	Auto	n
5R	J	$(2.5, 1.5)$		1	Estimated	Auto	n
6L	E	$(6, 5, 5)$		10	Estimated	Auto	n
6R	K	$(2.5, 1.5, 1.5)$		1	Estimated	Auto	n
7	L	No spike		1 10	Estimated	Auto	n

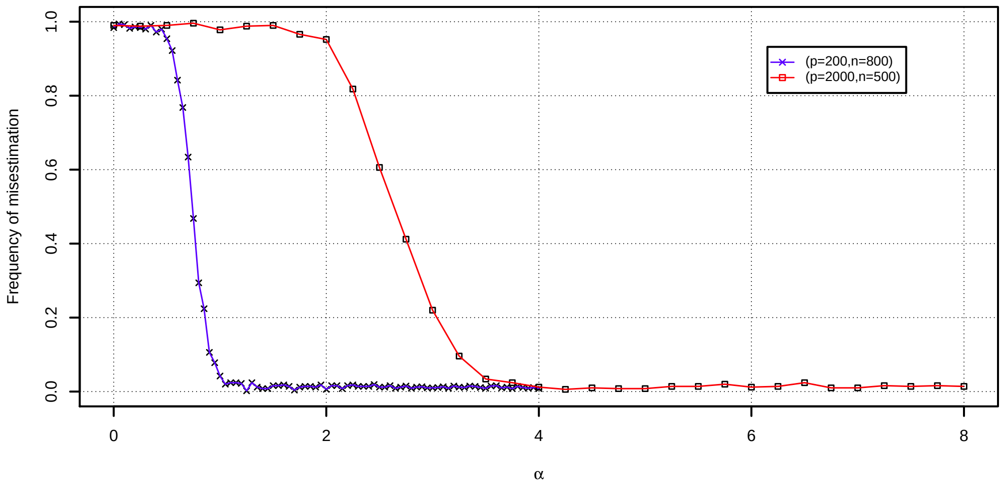


Figure 6.8 Misestimation rates as a function of spike strength for $(p, n) = (200, 800)$ and $(p, n) = (2000, 500)$.

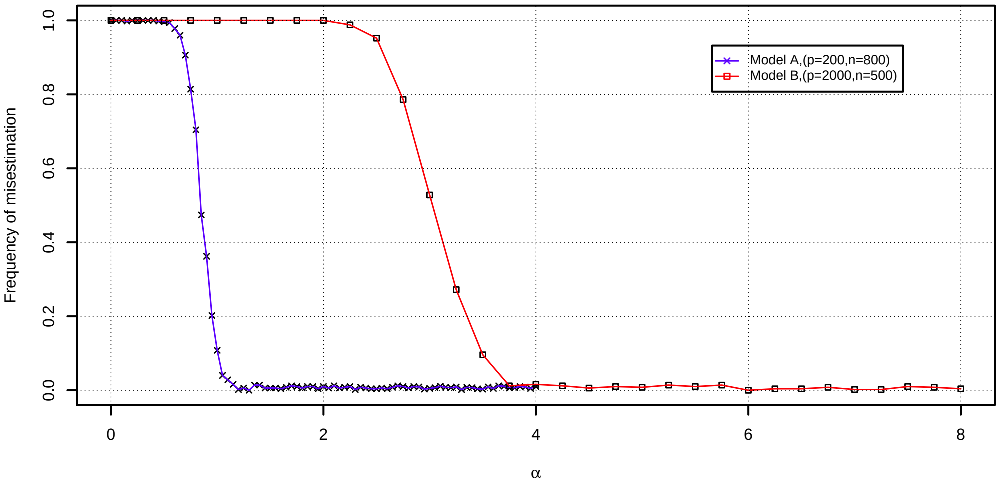


Figure 6.9 Misestimation rates as a function of spike strength for $(p, n) = (200, 800)$, Model A and $(p, n) = (2000, 500)$, Model B.

6.6.3 An automatic calibration procedure for the tuning parameter C

In the previous experiments, the tuning parameter C are selected manually on a case by case basis. This is however untenable in a real-life situation and an automatic calibration of this parameter is preferable. The idea is to use the difference of the two largest eigenvalues of a Wishart matrix (which correspond to the null case without any spike): indeed, the estimator $\hat{q}_n^*$ is found once two consecutive eigenvalues are below the threshold d_n corresponding to a noise eigenvalue. As the distribution of the difference between eigenvalues of a Wishart matrix is not known explicitly, 500 independent replications are drawn to evaluate numerically approximate the distribution of the difference between the two largest eigenvalues $\tilde{\lambda}_1 - \tilde{\lambda}_2$. The quantile s such that $\mathbb{P}(\tilde{\lambda}_1 - \tilde{\lambda}_2 \leq s) = 0.98$ is estimated by the average of the 10th and the 11th largest spacings. Finally, the automatically tuned value is set to

$$\tilde{C} = s \cdot n^{2/3} / \sqrt{2 \times \log\log(n)} \,. \tag{6.56}$$

The values of $\tilde{C}$ are reported in Table 6.2 for various (p, n) with $y = 1$ or $y = 10$.

Table 6.2 *Approximation of the threshold s such that* $\mathbb{P}(\tilde{\lambda}_1 - \tilde{\lambda}_2 \leq s) = 0.98$.

(p,n)	(200,200)	(400,400)	(600,600)	(2000,200)	(4000,400)	(7000,700)
Value of s	0.340	0.223	0.170	0.593	0.415	0.306
$\tilde{C}$	6.367	6.398	6.277	11.106	11.906	12.44

The values of $\tilde{C}$ are quite close to the manually tuned values found previously in similar settings (For instance, $C = 5$ for $y = 1$ and $C = 9.9$ or 11 for $y = 10$), although they are slightly higher. Therefore, this automatic calibration of $\tilde{C}$ can be used in practice for arbitrary pairs of (p, n).

To assess the quality of the automatic calibration procedure, some simulation experiments are run using both $\tilde{C}$ and the manually tuned C. The case $y = 10$ is considered in Figure 6.10. On the left panel, Model D ($\alpha = (6, 5)$) and Model E ($\alpha = (6, 5, 5)$) (upper curve) are considered, while the right panel reports on Model F ($\alpha = (10, 5)$) and Model G ($\alpha = (10, 5, 5)$) (upper curve). The dotted lines are the results with C manually tuned. Using the automatic value $\tilde{C}$ causes only a slight deterioration of the estimation performance. Notice however a significantly higher error rates in the case of multiple spikes for moderate sample sizes.

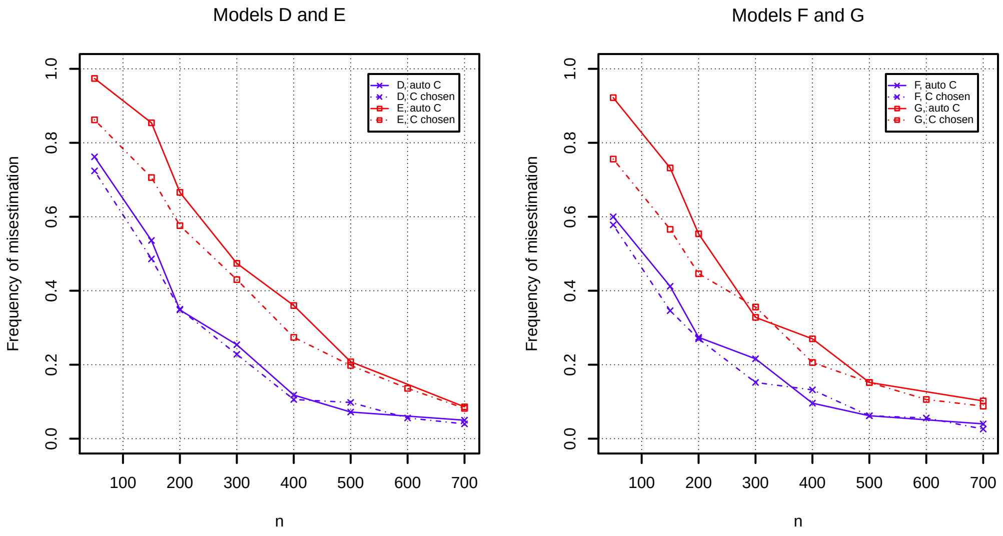


Figure 6.10 Misestimation rates as a function of n for Models D, E (left) and Models F, G (right).

The case $y = 1$ is considered in Figure 6.11 with Models H ($\alpha = 1.5$) and I ($\alpha = (1.5, 1.5)$) (upper curve) on the left and Model J ($\alpha = (2.5, 1.5)$) and K ($\alpha = (2.5, 1.5, 1.5)$) (upper curve) on the right.

Compared to the previous situation of $y = 10$, using the automatic value $\tilde{C}$ affects a bit more the estimator $\hat{q}_n^*$ (up to 20% of degradation). Nevertheless, the estimator remains consistent.

6.6.4 Method of Kritchman & Nadler and comparison

Detector of Kritchman & Nadler

A benchmark for the number of spikes with high-dimensional data is the *Kritchman and Nadler's detector* (KN). In this section, this detector is compared to $\hat{q}_n^*$ (denoted hereafter as PY) by simulations.

Recall that in the null case (without any spike) and assuming the variables $\{y_{ij}\}$ are

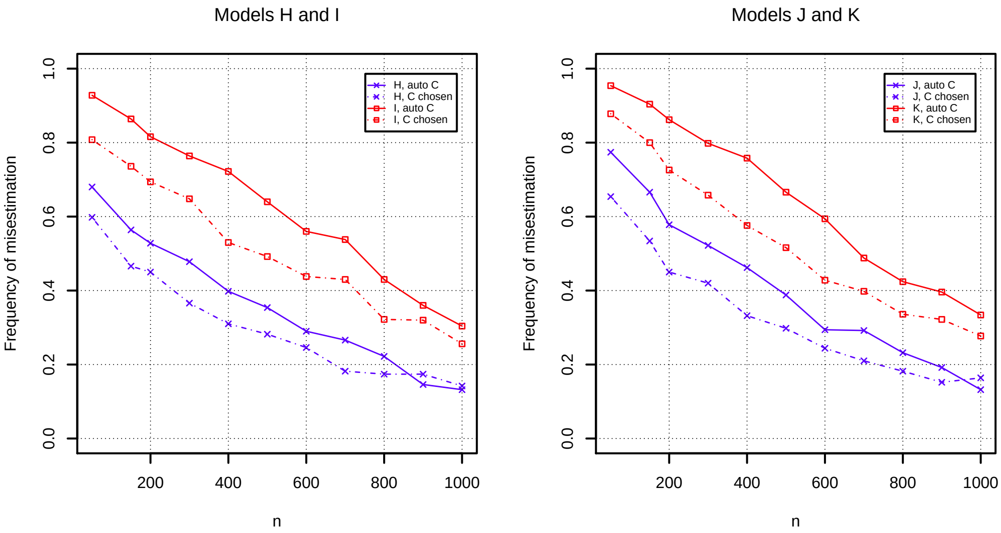


Figure 6.11 Misestimation rates as a function of n for Models H, I (left) and Models J, K (right).

Gaussian, the largest sample eigenvalue λ_1 obeys Tracy-Widom law (Theorem 6.3)

$$\mathbb{P}\left(\frac{\lambda_1}{\sigma^2} < \frac{\beta_p}{n^{2/3}}s + b\right) \to F_1(s), \qquad s > 0,$$

where $b = (1 + \sqrt{y})^2, \beta_p = \left(1 + \sqrt{\frac{p}{n}}\right)\left(1 + \sqrt{\frac{n}{p}}\right)^{\frac{1}{3}}$ and F_1 is the Tracy-Widom distribution of order 1. Assume that the variance σ^2 is known. To distinguish a spike eigenvalue λ from a noise one at an asymptotic significance level γ, the idea of the KN detector is to check whether

$$\lambda_k > \sigma^2\left(\frac{\beta_{p-k}}{n^{2/3}}s(\gamma) + b\right), \tag{6.57}$$

where $s(\gamma)$ verifies $F_1(s(\gamma)) = 1 - \gamma$ and can be found by inverting the Tracy-Widom distribution. The KN detector is based on a sequence of nested hypothesis tests of the following form: for $k = 1, 2, \ldots, \min(p, n) - 1$,

$$\mathcal{H}_0^{(k)}: m \leq k - 1 \quad vs. \quad \mathcal{H}_1^{(k)}: m \geq k\,.$$

For each value of k, if (6.57) is satisfied, $\mathcal{H}_0^{(k)}$ is rejected and k is increased by one. The procedure stops once an instance of $\mathcal{H}_0^{(k)}$ is accepted and the number of spikes is then estimated to be $\tilde{q}_n = k - 1$. Formally, their estimator is defined by

$$\tilde{q}_n = \underset{k}{\operatorname{argmin}}\left(\lambda_k < \widehat{\sigma}^2\left(\frac{\beta_{p-k}}{n^{2/3}}s(\gamma) + b\right)\right) - 1.$$

Here $\widehat{\sigma}$ is some estimator of the noise level σ^2.

Comparison between the KN and PY estimators

In order to follow a real-life situation, both estimators are run with an estimated noise variance $\hat{\sigma}^2$. Furthermore, the automatically calibrated value $\tilde{C}$ is used for the PY estimator. The value of $\gamma = 0.5\%$ is given to the false alarm rate of the estimator KN, as recommended by its authors.

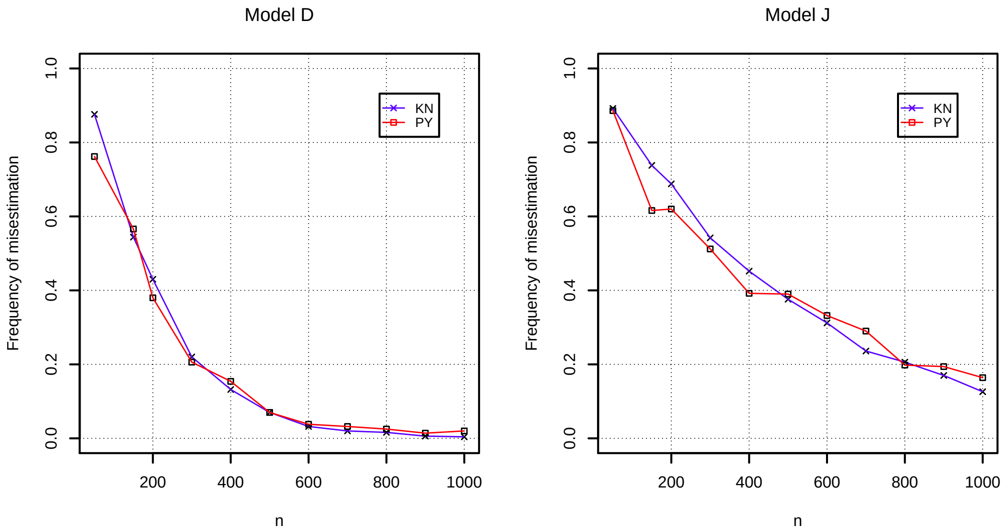


Figure 6.12 Misestimation rates as a function of n for Model D (left) and Model J (right).

In Figure 6.12, Model D: $(\alpha_1, \alpha_2) = (6, 5)$ and and Model J: $(\alpha_1, \alpha_2) = (2.5, 1.5)$ are considered. For both models, the performances of the two estimators are close. However the estimator PY is slightly better for moderate values of n ($n \leq 400$) while the estimator KN has a slightly better performance for larger n. The difference between the two estimators are more important for Model J (up to 5%).

Next in Figure 6.13 Model E: $(\alpha_1, \alpha_2, \alpha_2) = (6, 5, 5)$ and Model K: $(\alpha_1, \alpha_2, \alpha_2) = (2.5, 1.5, 1.5)$ are examined. These two models are analogous to Model D and J but with two multiple spikes.

For Model E, the estimator PY shows superior performance for $n \leq 500$ (up to 20% less error): adding a multiple spike affects more the performance of the estimator KN. The difference between the two algorithms for Model K is bigger than in the previous cases; the estimator PY performs better in all cases, up to 10%.

In Figure 6.14, the null case without any spike at all (Model L) is considered. The estimation rates become the so-called *false-alarm rate*, a concept widely used in signal processing literature. The cases of $y = 1$ and $y = 10$ with $\sigma^2 = 1$ given are considered. In both situations, the false-alarm rates of two estimators are quite low (less than 4%), and the detector KN has a lower false-alarm rate.

In summary, in most of situations reported here, the estimator $\hat{q}_n^*$ (PY) compares favourably to the benchmark KN detector. It is however important to notice a fundamental difference between these two estimators: the KN estimator is designed to keep the false alarm rate

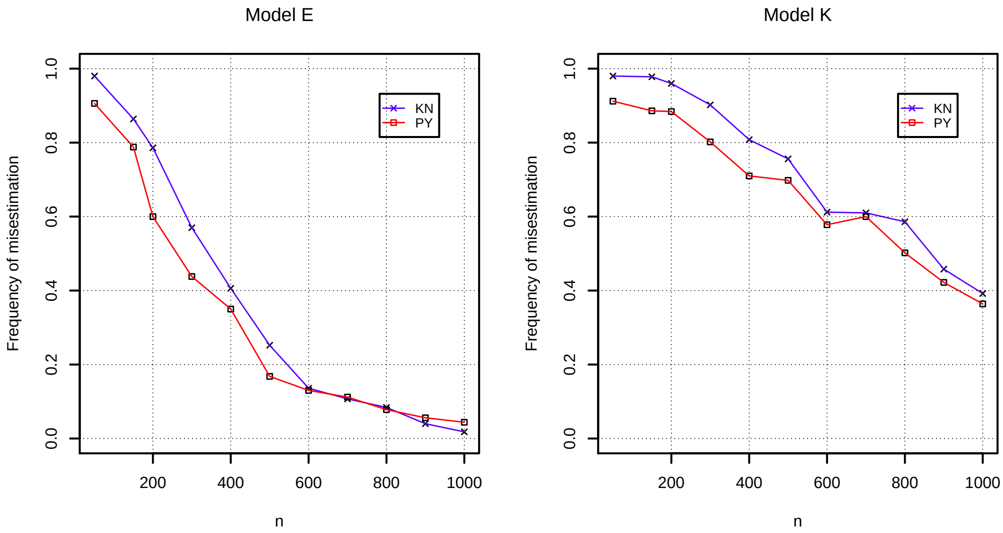


Figure 6.13 Misestimation rates as a function of n for Model E (left) and Model K (right).

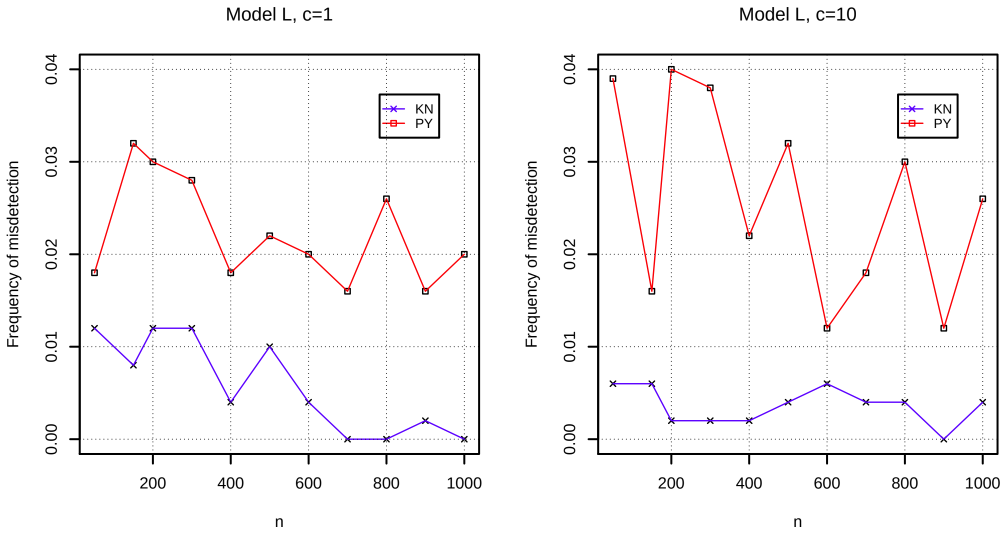


Figure 6.14 False-alarm rates as a function of n for $y = 1$ (left) and $y = 10$ (right).

at a very low level while the PY estimator attempts to minimise an overall misestimation rate.

6.7 Estimation of the noise variance

For Johnstone's spiked population model given in (6.50), the maximum likelihood estimator of the noise variance is given in (6.55), namely

$$\widehat{\sigma}^2 = \frac{1}{p-m} \sum_{i=m+1}^{p} \lambda_i, \tag{6.58}$$

i.e. the average the $p - m$ sample eigenvalues corresponding to the noise (Anderson and Rubin, 1956) (notice that this is derived under the Gaussian assumption). In the classical low-dimensional setting, we let p fixed and $n \to \infty$, asymptotic normality holds with the standard $\sqrt{n}$-convergence and we have (Anderson and Amemiya, 1988).

$$\sqrt{n}(\widehat{\sigma}^2 - \sigma^2) \xrightarrow{\mathcal{L}} \mathcal{N}(0, s^2), \quad s^2 = \frac{2\sigma^4}{p-m}. \tag{6.59}$$

Once again, the situation is radically different when p is large compared to the sample size n and it has been widely observed in the literature that $\widehat{\sigma}^2$ seriously underestimates the true noise variance σ^2 in such situation. As all meaningful inference procedures in the model will unavoidably use this variance estimate, such a severe bias is more than disappointing and needs to be corrected.

Notice that for the spiked population covariance matrix $\mathbf{\Sigma}$, its spectral distribution is

$$H_n = \frac{p-m}{p}\delta_{\sigma^2} + \frac{1}{p}\sum_{k=1}^{K} m_k \delta_{\alpha_i+\sigma^2}, \tag{6.60}$$

and $H_n \to \delta_{\sigma^2}$.

Theorem 6.19 *Assume that*

(a) Conditions (i)-(vi) on the spiked population model (6.50) *as formulated in §6.2 are satisfied and the variables $\{y_{ij}\}$ are Gaussian;*
(b) All the K spike eigenvalues are fundamental spikes.

Then, we have

$$\frac{(p-m)}{\sigma^2\sqrt{2y}}(\widehat{\sigma}^2 - \sigma^2) + b(\sigma^2) \xrightarrow{\mathscr{D}} \mathcal{N}(0,1),$$

where

$$b(\sigma^2) = \sqrt{\frac{y}{2}}\left(m + \sigma^2 \sum_{k=1}^{K} \frac{m_k}{\alpha_k}\right).$$

Proof By definition,

$$(p-m)\widehat{\sigma}^2 = \sum_{i=1}^{p} \lambda_i - \sum_{i=1}^{m} \lambda_i.$$

By Corollary 6.4,

$$\sum_{i=1}^{m} \lambda_i \xrightarrow{\text{a.s.}} \sum_{k=1}^{K} m_k \left(\alpha_k + \frac{c\sigma^4}{\alpha_k}\right) + \sigma^2 m(1+y). \tag{6.61}$$

For the first term, we have

$$\begin{aligned}\sum_{i=1}^{p}\lambda_i &= p\int xdF_n(x)\\ &= p\int x\,\mathrm{d}(F_n - F_{y_n,H_n})(x) + p\int x\,\mathrm{d}F_{y_n,H_n}(x)\\ &= X_n(f) + p\int x\,\mathrm{d}F_{y_n,H_n}(x),\end{aligned}$$

where F_n is the ESD of the sample covariance matrix $\mathbf{S}_n$ and the function f is the identity function $f(x) = x$. By Theorem 3.9, the first term is asymptotically normal

$$X_n(f) = \sum_{i=1}^{p}\lambda_i - p\int x\,\mathrm{d}F_{y_n,H_n}(x) \xrightarrow{\mathscr{D}} \mathcal{N}(m(x), v(x)),$$

with asymptotic mean

$$m(x) = 0\ , \tag{6.62}$$

and asymptotic variance

$$v(x) = 2c\sigma^4\ . \tag{6.63}$$

The derivation of these two formula and the following identity

$$\int x\,\mathrm{d}F_{y_n,H_n}(x) = \int t\,\mathrm{d}H_n(t) = \sigma^2 + \frac{1}{p}\sum_{k=1}^{K} m_k\alpha_i.$$

are standard and left to the reader. So we have

$$\sum_{i=1}^{p}\lambda_i - p\sigma^2 - \sum_{k=1}^{K} m_k\alpha_k \xrightarrow{\mathscr{D}} \mathcal{N}(0, 2y\sigma^4). \tag{6.64}$$

By (6.61) and (6.64) and using Slutsky's lemma, we obtain

$$(p-m)(\widehat{\sigma^2} - \sigma^2) + y\sigma^2\left(m + \sigma^2\sum_{k=1}^{K}\frac{m_k}{\alpha_k}\right) \xrightarrow{\mathscr{D}} \mathcal{N}(0, 2y\sigma^4).$$

□

Therefore for high-dimensional data, the m.l.e. $\widehat{\sigma^2}$ has an asymptotic bias $-b(\sigma^2)$ (after normalisation). This bias is a complex function of the noise variance and the m spiked eigenvalues. It is worth noticing that the above CLT is still valid if $\tilde{y}_n = (p-m)/n$ is substituted for y. Now if we let $p \ll n$ so that $\tilde{y}_n \simeq 0$ and $b(\sigma^2) \simeq 0$, and hence

$$\frac{(p-m)}{\sigma^2\sqrt{2y}}(\widehat{\sigma^2} - \sigma^2) + b(\sigma^2) \simeq \frac{\sqrt{p-m}}{\sigma^2\sqrt{2}}(\widehat{\sigma^2} - \sigma^2)\ .$$

This is nothing but the CLT (6.59) for $\widehat{\sigma^2}$ known under the classical low-dimensional scheme. From this point of view, Theorem 6.19 constitutes a natural extension of the classical CLT to the high-dimensional context.

6.7.1 Monte-Carlo experiments

In these experiments, i.i.d. Gaussian samples of size n are used in three different settings:

- Model 1: $\text{spec}(\mathbf{\Sigma}) = (25, 16, 9, 0, \dots, 0) + \sigma^2(1, \dots, 1)$, $\sigma^2 = 4$, $y = 1$;
- Model 2: $\text{spec}(\mathbf{\Sigma}) = (4, 3, 0, \dots, 0) + \sigma^2(1, \dots, 1)$, $\sigma^2 = 2$, $y = 0.2$;
- Model 3: $\text{spec}(\mathbf{\Sigma}) = (12, 10, 8, 8, 0, \dots, 0) + \sigma^2(1, \dots, 1)$, $\sigma^2 = 3$, $y = 1.5$.

Figure 6.15 presents the histograms from 1000 replications of

$$\frac{(p-m)}{\sigma^2\sqrt{2y}}(\widehat{\sigma}^2 - \sigma^2) + b(\sigma^2)$$

for the three models above, with different sample size n and $p = y \times n$, compared to the density of the standard normal distribution. Even for a moderate sample size like $n = 100$, the distribution is almost normal.

In Table 6.3, we compare the empirical bias of $\widehat{\sigma}^2$ (i.e. the empirical mean of $\sigma^2 - \widehat{\sigma}^2 = \sigma^2 - \frac{1}{p-m}\sum_{i=m+1}^{p}\lambda_i$) over 1000 replications with the theoretical one $-\sigma^2\sqrt{2y}b(\sigma^2)/(p-m)$ in different settings. In all the three models, the empirical and theoretical bias are close each other. As expected, their difference vanishes when p and n increase.

Table 6.3 *Comparison between the empirical and the theoretical bias in various settings.*

	Settings		Empirical bias	Theoretical bias	|Difference|
	$p = 100$	$n = 100$	-0.1556	-0.1589	0.0023
Model 1	$p = 400$	$n = 400$	-0.0379	-0.0388	0.0009
	$p = 800$	$n = 800$	-0.0189	-0.0193	0.0004
	$p = 20$	$n = 100$	-0.0654	-0.0704	0.0050
Model 2	$p = 80$	$n = 400$	-0.0150	-0.0162	0.0012
	$p = 200$	$n = 1000$	-0.0064	-0.0063	0.0001
	$p = 150$	$n = 100$	-0.0801	-0.0795	0.0006
Model 3	$p = 600$	$n = 400$	-0.0400	-0.0397	0.0003
	$p = 1500$	$n = 1000$	-0.0157	-0.0159	0.0002

6.7.2 A bias-corrected estimator

The previous theory recommends to correct the negative bias of $\widehat{\sigma}^2$. However, the bias $b(\sigma^2)$ depends on the number m and the values of the spikes α_k. These parameters could not be known in real-life applications and they need to be first estimated. Firstly, we can use the consistent estimators introduced in §6.6 for the unknown number m of spikes. Next, estimators presented in §6.5 will give consistent estimates of the values of the spikes.

As the bias depends also on σ^2 which we want to estimate, a natural correction is to use the plug-in estimator

$$\widehat{\sigma}^2_* = \widehat{\sigma}^2 + \frac{b(\widehat{\sigma}^2)}{p-m}\widehat{\sigma}^2\sqrt{2y}.$$

Notice that in this formula, the number of factors m can be replaces by any consistent

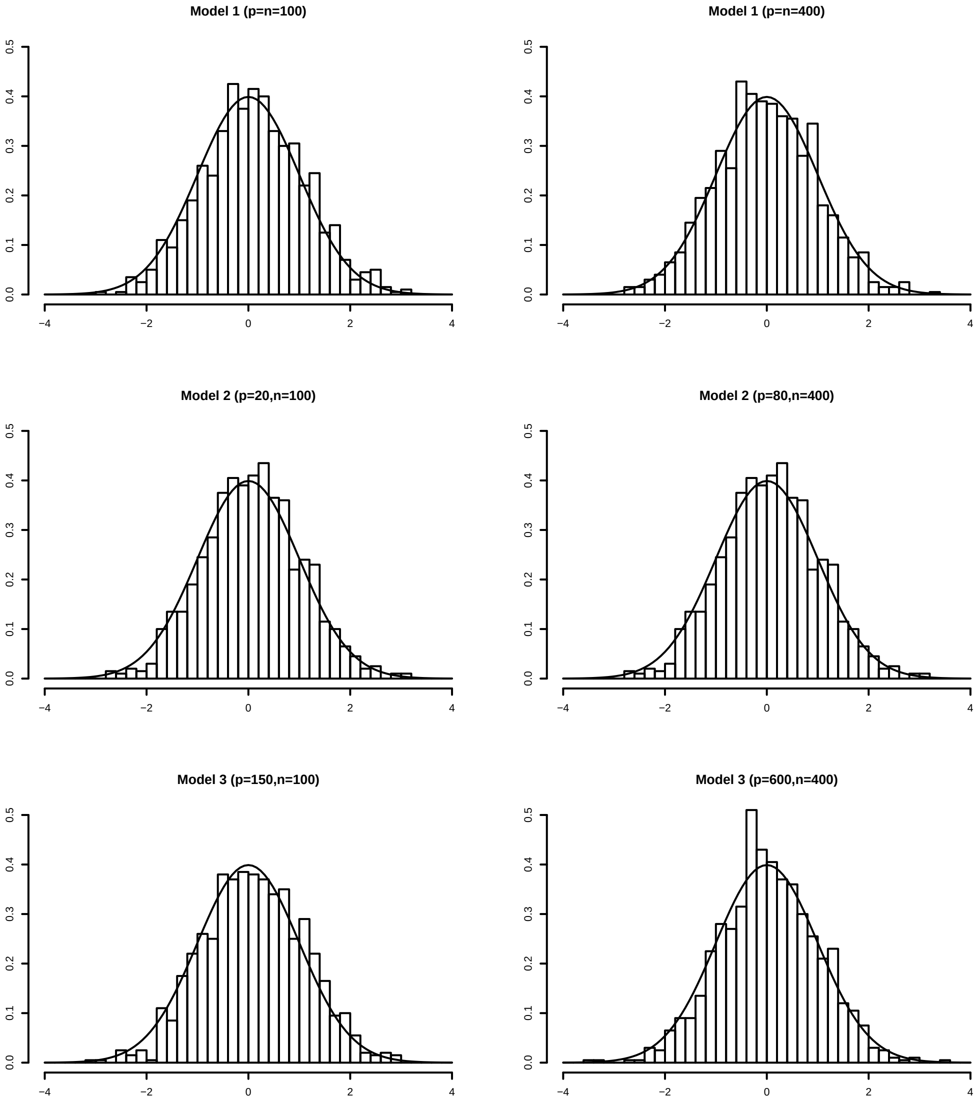


Figure 6.15 Histogram of $\frac{(p-m)}{\sigma^2\sqrt{2y}}(\widehat{\sigma}^2-\sigma^2)+b(\sigma^2)$ compared with the density of a standard Gaussian law.

estimate as discussed above without affecting its limiting distribution. Using Theorem 6.19 and the delta-method, we obtain the following CLT

Theorem 6.20 *We assume the same conditions as in Theorem 6.19. Then, we have*

$$\tilde{v}(y)^{-\frac{1}{2}}\left\{\widehat{\sigma}^2_* - \sigma^2 + \tilde{b}(\sigma^2)\right\} \xrightarrow{\mathscr{D}} \mathcal{N}(0,1),$$

where

$$\tilde{b}(\sigma^2) = \frac{y\sqrt{2c}\sigma^2}{(p-m)^2}\left(mb(\sigma^2) + 2\sigma^2 b(\sigma^2)\sum_{k=1}^{K} m_k\alpha_k^{-1}\right) - \frac{2c^2\sigma^4 b(\sigma^2)^2 \sum_{k=1}^{K} m_k\alpha_k^{-1}}{(p-m)^3} = O_p\left(\frac{1}{p^2}\right),$$

and

$$\tilde{v}(c) = \frac{2y\sigma^4}{(p-m)^2}\left(1 + \frac{ym}{p-m} + \frac{4c^2\sigma^4}{(p_m)^3}\sum_{k=1}^{K} m_k\alpha_i^{-1}\right)^2 = \frac{2y\sigma^4}{(p-m)^2}\left(1 + O_p\left(\frac{1}{p}\right)\right).$$

Basically, this theorem states that

$$\frac{p-m}{\sigma^2\sqrt{2y}}\left(\widehat{\sigma}_*^2 - \sigma^2\right) \xrightarrow{\mathscr{D}} \mathcal{N}(0,1).$$

Compared to the m.l.e. $\widehat{\sigma}^2$ in Theorem 6.19, the new estimator has no longer a bias after normalisation by $\frac{p-m}{\sigma^2\sqrt{2y}}$. The terms $\tilde{b}(\sigma^2)$ and $\tilde{v}(c)$ in the theorem give more details for the centring parameter and the normalisation rate.

To assess the quality of this bias-corrected estimator $\widehat{\sigma}_*^2$, we conduct some simulation experiments using the previous settings: Tables 6.4 and 6.5 give the empirical mean of $\widehat{\sigma}_*^2$ over 1000 replications compared with the empirical mean of $\widehat{\sigma}^2$, as well as the mean squared errors and mean absolute deviations. For comparison, the same statistics are also given for two alternative estimators $\widehat{\sigma}^2_{\mathrm{KN}}$ and $\widehat{\sigma}^2_{\mathrm{US}}$ defined as follow:

- $\widehat{\sigma}^2_{\mathrm{KN}}$ is the solution of the following non-linear system of $m+1$ equations involving the $m+1$ unknowns $\widehat{\rho}_1, \ldots, \widehat{\rho}_m$ and $\widehat{\sigma}^2_{\mathrm{KN}}$

$$\widehat{\sigma}^2_{\mathrm{KN}} - \frac{1}{p-m}\left[\sum_{j=m+1}^{p} \lambda_{n,j} + \sum_{j=1}^{m}(\lambda_{n,j} - \widehat{\rho}_j)\right] = 0,$$

$$\widehat{\rho}_j^2 - \widehat{\rho}_j\left(\lambda_{n,j} + \widehat{\sigma}^2_{\mathrm{KN}} - \widehat{\sigma}^2_{\mathrm{KN}}\frac{p-m}{n}\right) + \lambda_{n,j}\widehat{\sigma}^2_{\mathrm{KN}} = 0.$$

- $\widehat{\sigma}^2_{\mathrm{US}}$ is defined as

$$\widehat{\sigma}^2_{\mathrm{US}} = \frac{\mathrm{median}(\lambda_{n,m+1}, \ldots, \lambda_{n,p})}{p_y^{-1}(0.5)},$$

where p_y^{-1} is quantile function of the Marčenko-Pastur distribution F_y.

In all three models considered, the bias-corrected estimator $\widehat{\sigma}_*^2$ is far much better than the original m.l.e. $\widehat{\sigma}^2$: here mean absolute deviations are reduced by 95% at least. The performances of $\widehat{\sigma}_*^2$ and $\widehat{\sigma}^2_{\mathrm{KN}}$ are similar. The estimator $\widehat{\sigma^2}_{\mathrm{US}}$ shows slightly better performance than the m.l.e. $\widehat{\sigma}^2$, but performs poorly compared to $\widehat{\sigma}_*^2$ and $\widehat{\sigma}^2_{\mathrm{KN}}$. Notice however the theoretic properties of $\widehat{\sigma}^2_{\mathrm{KN}}$ and $\widehat{\sigma^2}_{\mathrm{US}}$ are unknown and so far there have been checked via simulations only.

Notes

The phoneme example discussed in §6.1 appeared in Johnstone (2001) and originated from Buja et al. (1995). The data set can be downloaded from

Table 6.4 *Empirical mean, MSE (between brackets) and mean absolute deviation of $\widehat{\sigma}^2$ and $\widehat{\sigma}^2_*$ in various settings.*

	Settings						
Mod.	p	n	σ^2	$\widehat{\sigma}^2$	$\lvert\sigma^2 - \widehat{\sigma}^2\rvert$	$\widehat{\sigma}^2_*$	$\lvert\sigma^2 - \widehat{\sigma}^2_*\rvert$
1	100	100	4	3.8464 (0.0032)	0.1536	3.9979 (0.0035)	0.0021
	400	400		3.9616 (0.0002)	0.0384	4.0000 (0.0002)	$< 10^{-5}$
	800	800		3.9809 (0.0001)	0.0191	4.0002 (0.0001)	0.0002
2	20	100	2	1.9340 (0.0043)	0.0660	2.0012 (0.0047)	0.0012
	80	400		1.9841 (0.0003)	0.0159	2.0001 (0.0003)	0.0001
	200	1000		1.9939 ($< 10^{-5}$)	0.0061	2.0002 ($< 10^{-5}$)	0.0002
3	150	100	3	2.8400 (0.0011)	0.1600	2.9926 (0.0013)	0.0074
	600	400		2.9605 (0.0001)	0.0395	2.9999 (0.0001)	0.0001
	1500	1000		2.9839 ($< 10^{-5}$)	0.0161	2.9998 ($< 10^{-5}$)	0.0002

Table 6.5 *Empirical mean, MSE (between brackets) and mean absolute deviation of $\widehat{\sigma}^2_{KN}$ and $\widehat{\sigma}^2_{US}$ in various settings.*

	Settings						
Mod.	p	n	σ^2	$\widehat{\sigma}^2_{\mathrm{KN}}$	$\lvert\sigma^2 - \widehat{\sigma}^2_{\mathrm{KN}}\rvert$	$\widehat{\sigma}^2_{\mathrm{US}}$	$\lvert\sigma^2 - \widehat{\sigma}^2_{\mathrm{US}}\rvert$
1	100	100	4	4.0030 (0.0036)	0.0030	3.8384 (0.0154)	0.1616
	400	400		4.0003 (0.0002)	0.0003	3.9585 (0.0013)	0.0415
	800	800		4.0002 (0.0001)	0.0002	3.9794 (0.0004)	0.0206
2	20	100	2	1.9997 (0.0048)	0.0003	1.9400 (0.0087)	0.0600
	80	400		2.0001 (0.0003)	0.0001	1.9851 (0.0008)	0.0149
	200	1000		2.0002 ($< 10^{-5}$)	0.0002	1.9942 (0.0001)	0.0058
3	150	100	3	2.9935 (0.0016)	0.0065	2.7750 (0.0092)	0.2250
	600	400		3.0006 (0.0001)	0.0006	2.9450 (0.0007)	0.0550
	1500	1000		2.9999 ($< 10^{-5}$)	0.0001	2.9773 (0.0001)	0.0227

`http://statweb.stanford.edu/%7Etibs/ElemStatLearn/`

the website of Hastie et al. (2009) (`Data` tab, then `Phoneme` entry). Figure 6.1 is produced using the first 162 observations in the section "dcl" of the data set.

The name of spiked population model is coined in Johnstone (2001) whilst the main purpose of the paper is the establishment of the Trace-Widom law (6.3) in the null case.

For Johnstone's spiked population model, the fluctuation of largest sample eigenvalues λ_j from a complex Gaussian population with a spiked covariance matrix is studied in Baik et al. (2005). These authors prove a transition phenomenon: the weak limit and the scaling of λ_j are different according to the location of underlying population spike eigenvalues with respect to the critical value $1 + \sqrt{y}$. In Baik and Silverstein (2006), the authors consider the spiked population model with general random variables: complex or real and not necessarily Gaussian. For the almost sure limits of the extreme sample eigenvalues, they also find that these limits depend on the critical values $1 + \sqrt{y}$ for largest sample eigenvalues, and on $1 - \sqrt{y}$ for smallest ones. In Paul (2007), a CLT is established for spiked sample eigenvalues under the Gaussian assumption and assuming that spikes are simple (multiplicity 1). The CLT for spiked sample eigenvalues in general case with general entries and arbitrary multiplicity numbers of the spikes is given in Bai and Yao (2008) (with limits located outside the Marčenko-Pastur bulk spectrum interval $[(1-\sqrt{y})^2, (1+\sqrt{y})^2]$.

The *generalised spiked population model* in §6.2 is due to Bai and Yao (2012) as well as most of the results of the section. The central limit theory in §6.4 follows Bai and Yao (2008).

For inference on a spiked population model, results in §6.5.2 are due to Bai and Ding (2012). This reference contains also a CLT for the estimator of spike eigenvalues. As for estimation of the number of spikes, the benchmark Kritchman and Nadler detector is introduced in Kritchman (2008) and Kritchman and Nadler (2009). In these papers, this detector is compared with other existing estimators in the signal processing literature, based on the minimum description length (MDL), Bayesian information criterion (BIC) and Akaike information criterion (AIC) (Wax and Kailath, 1985). In most of the studied cases, the Kritchman and Nadler estimator performs better in case of high-dimensional data. Furthermore in Nadler (2010), this estimator is also compared with an improved AIC estimator and it still has a better performance. Therefore, in this chapter comparison for the PY estimator $\hat{q}_n^*$ is only made with the above benchmark detector. The presentation of §6.6 follows Passemier and Yao (2014) which generalises a previous work Passemier and Yao (2012) by the same authors. Finally, the material in §6.7 is borrowed from Passemier et al. (2017). Notice that the two alternative estimators given there are due to Kritchman (2008) and Ulfarsson and Solo (2008), respectively.

The spiked population model is closely connected to other random matrices ensembles through the general concept of small-rank perturbations. The goal is again to examine the effect caused on the sample extreme eigenvalues by such perturbations. Theories on perturbed Wigner matrices can be found in Péché (2006), Féral and Péché (2007), Capitaine et al. (2009), Pizzo et al. (2013) and Renfrew and Soshnikov (2013). In a more general setting of finite-rank perturbation including both the additive and the multiplicative one, point-wisely convergence of extreme eigenvalues is established in Benaych-Georges and Nadakuditi (2011) while their fluctuations are studied in Benaych-Georges et al. (2011). In addition, Benaych-Georges and Nadakuditi (2011) contain also results on spiked eigenvectors that are similar to those presented in §6.3.

Appendix A

Curvilinear integrals

This appendix gives a short introduction to the theory of curvilinear and contour integrals in the complex plane. As in the CLT's developed in Chapter 3 for linear spectral statistics of sample covariance matrices and of random Fisher matrices, the mean and covariance functions of the limiting Gaussian distributions are expressed in terms of contour integrals, explicit calculations of such contour integrals frequently appear in various chapters of this book. This appendix will thus provide useful and self-contained references for those calculations.

The presentation here follows the lectures notes in complex analysis given by André Giroux at Université de Montréal (Giroux, 2013). In particular, interested reader is recommended to consult this reference for detailed proofs of the results introduced in this chapter.

§1 Let f be a complex-valued function defined on an open subset $D \subseteq \mathbb{C}$ of the complex plane and $z_0 \in D$. The function f is *differentiable at* z_0 if

$$\lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}$$

exists. In this is the case, the limit is simply denoted by $f'(z_0)$.

The function f is *holomorphic* in D if it is differentiable everywhere in D. A function is said *holomorphic at a point* if it is holomorphic in an open disk centred at this point.

The term *analytic function* is often used interchangeably with holomorphic function, although the word *analytic* is also used in a broader sense to describe any function (real, complex, or of more general type) that can be written as a convergent power series in a neighborhood of each point in its domain. The fact that the class of complex analytic functions coincides with the class of holomorphic functions is a major theorem in complex analysis.

§2 **Theorem A.1** *(Cauchy-Riemann) Let $D \subseteq \mathbb{C}$ and $f : D \to \mathbb{C}$ be a holomorphic function. Then the partial derivatives of the real and imaginary parts u and v of f exist everywhere in D and they satisfy the Cauchy-Riemann equations*

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}.$$

§3 A *differentiable curve* $\mathcal{C}$ is defined by a function $t \mapsto z(t)$ from $[a, b] \subseteq \mathbb{R}$ to $\mathbb{C}$ which is differentiable with a continuous and non-null derivative $z'(t) = x'(t) + iy'(t)$:

$$\mathcal{C} = \{z \mid z = z(t),\ a \leq t \leq b\}\,.$$

The function z is called a *parametrisation* of the curve. A *piece-wisely differentiable curve*, or a *path*, is obtained by joining a finite number of differentiable curves.

Let $\mathcal{C}$ be a path. It will always be assumed that $z(t_1) \neq z(t_2)$ for $a < t_1 < t_2 < b$, and when $z(a) = z(b)$, the path is said *closed*. A closed path partition the plane in two disjoint domains: one is bounded and called *the interior* of $\mathcal{C}$, the other is unbounded and called *the exterior* of $\mathcal{C}$ (Jordan theorem).

A closed path is also called a *contour*.

Example A.2 The unit circle is a contour and can be parametrised by $z(t) = e^{it}$, $0 \leq t \leq 2\pi$. Its interior is the (open) unit disk $D(0, 1)$.

A parametrisation of a closed curve implies a *running direction* of the curve. A closed curve is *anticlockwise run* if the vector $iz'(t)$ points in the direction of its interior; otherwise it is *clockwise run*.

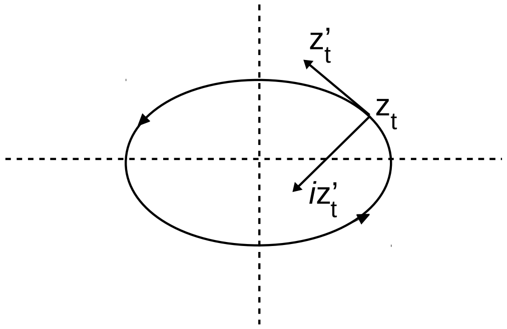


Figure A.1 An anticlockwise-run closed curve.

§4 A subset $D \subseteq \mathbb{C}$ which is open and connected is called a *domain*. Let $D \subseteq \mathbb{C}$ be a domain, $f : D \longrightarrow \mathbb{C}$ a continuous function and $\mathcal{C}$ a differentiable curve with parametrisation $z = z(t),\ a \leq t \leq b$. The formula

$$\int_{\mathcal{C}} f(z)dz = \int_a^b f(z(t))z'(t)dt\,,$$

defines the *curvilinear integral* of f along the curve $\mathcal{C}$. When $\mathcal{C}$ is a contour, the curvilinear integral becomes a *contour integral*, and we use a special notation $\oint_{\mathcal{C}} f(z)dz$ for such integral.

It is easily checked that the value of a curvilinear integral is independent of the choice of parametrisation. The curvilinear integral along a path $\mathcal{C}_1 + \mathcal{C}_2$ is defined by

$$\int_{\mathcal{C}_1+\mathcal{C}_2} f(z)dz = \int_{\mathcal{C}_1} f(z)dz + \int_{\mathcal{C}_2} f(z)dz.$$

If $f = F'$ has a primitive function F which is holomorphic on D, then

$$\int_{\mathcal{C}} f(z)dz = \int_a^b f(z(t))z'(t)dt$$
$$= \int_a^b F'(z(t))z'(t)dt = F(z(t))|_a^b = F(z_2) - F(z_1).$$

It is thus reasonable to write

$$\int_{z_1}^{z_2} f(z)dz = F(z_2) - F(z_1).$$

In particular for a contour $\mathcal{C}$,

$$\oint_{\mathcal{C}} f(z)dz = 0\ .$$

Example A.3 Let

$$\mathcal{C}_1 = \left\{z \mid e^{it},\ 0 \le t \le \pi\right\},$$

and

$$\mathcal{C}_2 = \left\{z \mid e^{-it},\ 0 \le t \le \pi\right\}.$$

Then

$$\int_{\mathcal{C}_1} \frac{dz}{z} = \int_0^{\pi} idt = i\pi,$$

and

$$\int_{\mathcal{C}_2} \frac{dz}{z} = \int_0^{\pi} -idt = -i\pi.$$

The curve $\mathcal{C}_1 - \mathcal{C}_2$, i.e. $\mathcal{C}_1 + (-\mathcal{C}_2)$, is the unit circle run anticlockwise, and we have

$$\int_{\mathcal{C}_1-\mathcal{C}_2} \frac{dz}{z} = 2i\pi.$$

The holomorphic function $1/z$ thus has no holomorphic primitive function in the origin-emptied complex plan.

§5 **Theorem A.4** *(Cauchy) Let $D \subseteq \mathbb{C}$ be a domain, $f : D \to \mathbb{C}$ a holomorphic function, and $\mathcal{C}$ a contour included in D together with its interior. Then*

$$\oint_{\mathcal{C}} f(z)dz = 0.$$

Theorem A.5 *(Cauchy) Let $D \subseteq \mathbb{C}$ be a domain, $f : D \to \mathbb{C}$ a holomorphic function, and $\mathcal{C}$ a contour included in D together with its interior. Then for any z in the interior of $\mathcal{C}$,*

$$f(z) = \frac{1}{2\pi i} \oint_{\mathcal{C}} \frac{f(\zeta)}{\zeta - z} d\zeta,$$

where the contour $\mathcal{C}$ is run anticlockwise.

Theorem A.6 *(Cauchy) Let $D \subseteq \mathbb{C}$ be a domain, $f : D \to \mathbb{C}$ a holomorphic function. Then, its derivative $f' : D \to \mathbb{C}$ is holomorphic. Moreover, on any contour $\mathcal{C}$ included in D together with its interior,*

$$f'(z) = \frac{1}{2\pi i} \oint_{\mathcal{C}} \frac{f(\zeta)}{(\zeta - z)^2} d\zeta,$$

where the contour $\mathcal{C}$ is run anticlockwise.

By iterating the differentiation, we see that under the same conditions as in the last theorem, all the derivatives of f are holomorphic and they satisfy the identity

$$f^{(n)}(z) = \frac{n!}{2\pi i} \oint_{\mathcal{C}} \frac{f(\zeta)}{(\zeta - z)^{n+1}} d\zeta,$$

for any integer $n \geq 1$. In other words, we can differentiate under the integral as often as we desire.

§6 A domain $D \subseteq \mathbb{C}$ is *simply connected* if each closed curve in D has its interior also included in D. For example, convex sets are simply connected, but annuli are not.

Theorem A.7 *A holomorphic function f on a simply connected domain has a holomorphic primitive function F on this domain:*

$$F(z) = F(z_0) + \int_{z_0}^{z} f(z) dz,$$

where $z_0 \in D$ and $F(z_0)$ are arbitrary.

Example A.8 The arctan function has a holomorphic extension on $\mathbb{C} \setminus \{(-i\infty, -i] \cup [i, +i\infty)\}$ defined by

$$\arctan z = \int_0^z \frac{d\zeta}{1 + \zeta^2}.$$

Therefore,

$$\begin{aligned}\arctan(1 + i) &= \int_0^1 \frac{1 + i}{1 + (1 + i)^2 t^2} dt \\ &= \int_0^1 \frac{1 + 2t^2}{1 + 4t^4} dt + i \int_0^1 \frac{1 - 2t^2}{1 + 4t^4} dt = 1.017 + i0.402 .\end{aligned}$$

§7 **Theorem A.9** *(Laurent) Let $D \subseteq \mathbb{C}$ be a domain including the annulus $\{z \mid r \leq |z - z_0| \leq R\}$ and $f : D \to \mathbb{C}$ a holomorphic function in D. Then*

$$f(z) = \sum_{k=-\infty}^{\infty} a_k (z - z_0)^k, \quad r < |z - z_0| < R ,$$

where

$$a_k = \frac{1}{2\pi i} \oint_{C_\rho} \frac{f(\zeta)}{(\zeta - z_0)^{k+1}} d\zeta,$$

and C_ρ is the circle centred at z_0 with radius ρ ($r < \rho < R$) and anticlockwise run.

The series in the theorem is called *Laurent series* of f at z_0. A plot of annulus where this expansion takes place is given on Figure A.2.

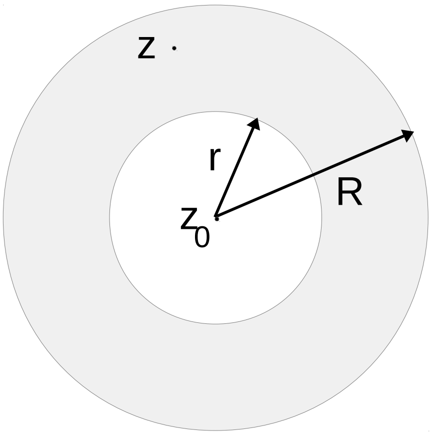


Figure A.2 Laurent Theorem.

§8 A point z_0 is an *isolated singularity* of a function f if the function is holomorphic in a punctured disk $\{z \mid 0 < |z - z_0| \leq R\}$ centred at z_0. According to the nature of the Laurent series of f, three types of singularity points exist.

- A *removable singularity point* when the Laurent series at the point has no terms with negative power $k < 0$;
- A *pole of order m* is an isolated singularity point such that

$$f(z) = \sum_{k=-m}^{+\infty} a_k(z - z_0)^k, \quad a_{-m} \neq 0 \; ;$$

- An *essential singularity point* is an isolated singularity point where the Laurent series has infinitely many terms with negative power $k < 0$.

In the neighbourhood of a pole z_0, a holomorphic function $f(z)$ tends to infinity when $z \to z_0$. The behaviour near an essential singularity is much more complex.

Example A.10 The function $e^{1/z}$ has an essential singularity at the origin. We have

$$\lim_{x\to 0_+} e^{1/x} = +\infty \; ,$$
$$\lim_{x\to 0_-} e^{1/x} = 0 \; ,$$
$$\lim_{y\to 0} e^{1/(iy)} \text{ does not exist.}$$

§9 A *meromorphic function* on a domain D is a function $f : D \to \mathbb{C}$ which is holomorphic on D except at isolated singularity points which are all poles. By letting the value of the function be ∞ at these poles, the function can be considered as a continuous function from D to $\overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$.

Example A.11 A rational function as well the following functions are meromorphic on the whole complex plan:

$$\frac{\sin z}{z^3}, \quad \frac{\cos z}{z^3 - 1}, \quad \text{and} \quad \tan z = \frac{\sin z}{\cos z}.$$

§10 Consider a function f holomorphic on the punctured disk $0 < |z - z_0| < r$ and let

$$f(z) = \sum_{k=-\infty}^{+\infty} a_k(z - z_0)^k, \quad 0 < |z - z_0| < r,$$

be its Laurent series at z_0. The *residue* of f at z_0 is

$$\operatorname{Res}(f, z_0) = a_{-1} = \frac{1}{2\pi i} \oint_{C_\rho} \frac{f(\zeta)}{\zeta - z_0} d\zeta, \tag{A.1}$$

where C_ρ is the circle centred at z_0 with radius $0 < \rho < r$ and run anticlockwise.

When z_0 is a pole of order m, the residue is also given by the formula

$$\operatorname{Res}(f, z_0) = \lim_{z \to z_0} \frac{1}{(m-1)!} \frac{d^{m-1}}{dz^{m-1}}[(z - z_0)^m f(z)]. \tag{A.2}$$

Example A.12 (a). $\operatorname{Res}\left(\frac{\sin z}{z^3}, 0\right) = 0.$ Indeed, $\sin z$ is holomorphic in $\mathbb{C}$ so that z is a pole of order 3 of the function $\sin z/z^3$. By (A.2),

$$\operatorname{Res}\left(\frac{\sin z}{z^3}, 0\right) = \lim_{z \to 0} \frac{1}{2!} \frac{d^2}{dz^2}[\sin(z)] = 0.$$

(b). $\operatorname{Res}\left(\frac{\cos z}{z^n - 1}, \omega_n^k\right) = \frac{1}{n}\omega_n^k \cos \omega_n^k, \quad \omega_n = e^{i2\pi/n}, \quad 0 \le k \le n-1.$ By definition,

$$z^n - 1 = \prod_{\ell=0}^{n-1}(z - \omega_n^\ell).$$

As $\cos z$ is holomorphic, ω_n^k is a simple pole of $\cos z/(z^n - 1)$. By (A.2),

$$\operatorname{Res}\left(\frac{\cos z}{z^n - 1}, \omega_n^k\right) = \frac{\cos \omega_n^k}{\prod_{j \neq k}(\omega_n^k - \omega_n^j)}.$$

On the other hand, by differentiation of $z^n - 1$, we have

$$nz^{n-1} = \sum_{\ell=0}^{n-1} \prod_{j \neq \ell}(z - \omega_n^j).$$

In particular, for $z = \omega_n^k$,

$$n(\omega_n^k)^{n-1} = \sum_{\ell=0}^{n-1} \prod_{j \neq \ell}(\omega_n^k - \omega_n^j) = \prod_{j \neq k}(\omega_n^k - \omega_n^j).$$

Therefore, since $(\omega_n^k)^n = 1$, we have

$$\operatorname{Res}\left(\frac{\cos z}{z^n - 1}, \omega_n^k\right) = \frac{\cos \omega_n^k}{n\omega_n^{-k}} = \frac{1}{n}\omega_n^k \cos \omega_n^k.$$

(c). $\operatorname{Res}\left(e^{1/z}, 0\right) = 1.$ As noticed previously, 0 is not a pole but an essential singularity of $e^{1/z}$. By (A.1),

$$\operatorname{Res}\left(e^{1/z}, 0\right) = \frac{1}{2\pi i} \oint_{C_\rho} \frac{e^{1/z}}{z} dz.$$

Using the parametrisation $z = \rho e^{i\theta}$ $(0 \leq \theta \leq 2\pi)$ of C_ρ, we have

$$\begin{aligned}\operatorname{Res}\left(e^{1/z}, 0\right) &= \frac{1}{2\pi}\int_0^{2\pi} \exp\{\rho^{-1}e^{-i\theta}\}d\theta \\ &= \sum_{k=0}^{\infty}\frac{1}{2\pi}\int_0^{2\pi}\frac{1}{k!}\rho^{-k}e^{-ik\theta}d\theta \\ &= \frac{1}{2\pi}\int_0^{2\pi} d\theta = 1.\end{aligned}$$

§11 **Theorem A.13** *Let $D \subseteq \mathbb{C}$ be a domain and $f : D \to \mathbb{C}$ a holomorphic function in D except at isolated singularity points. Let $\mathcal{C}$ be a contour which is included in D together with its interior, does not pass by any of these singularity points and contain a finite number of them, say $z_1, \ldots, z_m$, in its interior. Then,*

$$\oint_{\mathcal{C}} f(z)dz = 2\pi i \sum_{k=1}^{n} Res(f, z_k), \tag{A.3}$$

where the contour $\mathcal{C}$ is run anticlockwise.

This theorem provides the fundamental tool used throughout the book for the calculation of useful contour integrals by the so-called *method of residue*.

Notice that the Cauchy formula

$$f^{(n)}(z_0) = \frac{n!}{2\pi i}\oint_{\mathcal{C}} \frac{f(z)}{(z - z_0)^{n+1}}dz,$$

corresponds to the case of a pole of order $n + 1$ at z_0.

Example A.14 We have

$$\oint_{\mathcal{C}} \frac{\sin z}{z^3} = 0\ ;$$

$$\oint_{\mathcal{C}} \frac{\cos z}{z^n - 1}dz = \frac{2\pi i}{n}\sum_k \omega_n^k \cos\omega_n^k,$$ where the sum runs over the ω_n^k's in the interior of $\mathcal{C}$;

$$\oint_{\mathcal{C}} e^{1/z}dz = \begin{cases} 1, & \text{if 0 is interior to } \mathcal{C}, \\ 0, & \text{otherwise.}\end{cases}$$

Appendix B

Eigenvalue inequalities

In this appendix, we list a series of inequalities on eigenvalues or singular values of complex-valued matrices that are used at several places of the book.

If $\mathbf{A}$ is a $p{\times}n$ matrix of complex entries, then its singular values $s_1(\mathbf{A}) \geq ... \geq s_q(\mathbf{A}) \geq 0$, $q = \min(p, n)$, are defined as the square roots of the q largest eigenvalues of the nonnegative definite Hermitian matrix $\mathbf{A}\mathbf{A}^*$ where $*$ denotes transpose and conjugate. If $\mathbf{A}$ $(n \times n)$ is Hermitian, then let $\lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq ... \geq \lambda_n(\mathbf{A})$ denote its eigenvalues.

Theorem B.1 *(singular value and spectral decomposition) Let $\mathbf{A}$ be a $p{\times}n$ matrix. Then there exist q orthonormal vectors $\boldsymbol{u}_1, ..., \boldsymbol{u}_q$ of $\mathbb{C}^p$ and q orthonormal vectors $\boldsymbol{v}_1, ..., \boldsymbol{v}_q$ of $\mathbb{C}^n$ such that*

$$\mathbf{A} = \sum_{j=1}^{q} s_j \boldsymbol{u}_j \boldsymbol{v}_j^* \tag{B.1}$$

From this expression, we immediately get the well-known Courant-Fischer fromula

$$s_k = \min_{w_1,...,w_{k-1}} \max_{\substack{\|\boldsymbol{v}\|_2=1 \\ \boldsymbol{v}\perp w_1,...,w_{k-1}}} \|A\boldsymbol{v}\|_2 \,. \tag{B.2}$$

If $\mathbf{A}$ is an $n \times n$ Hermitian matrix,then there exist an orthonormal basis $\{\boldsymbol{u}_1, ..., \boldsymbol{u}_n\}$ of $\mathbb{C}^n$ such that

$$A = \sum_{j=1}^{n} \lambda_j \boldsymbol{u}_j \boldsymbol{u}_j^* \tag{B.3}$$

Similarly, we have the formula

$$\lambda_k = \min_{w_1,...,w_{k-1}} \max_{\substack{\|\boldsymbol{v}\|_2=1 \\ \boldsymbol{v}\perp w_1,...,w_{k-1}}} \boldsymbol{v}^* A \boldsymbol{v}. \tag{B.4}$$

Theorem B.2 *(Cauchy interlacing law) For any $n \times n$ Hermitian matrix $\mathbf{A}_n$ with top left $(n-1)\times(n-1)$ minor $\mathbf{A}_{n-1}$, then*

$$\lambda_{i+1}(\mathbf{A}_n) \leq \lambda_i(\mathbf{A}_{n-1}) \leq \lambda_i(\mathbf{A}_n), \tag{B.5}$$

for all $1 \leq i < n$.

Notice that if one takes successive minors $\mathbf{A}_{n-1}, \mathbf{A}_{n-2}, \ldots, \mathbf{A}_1$ of an $n{\times}n$ Hermitian matrix $\mathbf{A}_n$, and computes their spectra, then (B.5) shows that this triangular array of numbers forms a pattern known as a *Gelfand-Tsetlin pattern*.

Theorem B.3 *(Weyl inequalities) For $n \times n$ Hermitian matrices* $\mathbf{A}$ *and* $\mathbf{B}$*,*

$$\lambda_{i+j-1}(\mathbf{A}+\mathbf{B}) \leq \lambda_i(\mathbf{A}) + \lambda_j(\mathbf{B}), \tag{B.6}$$

valid whenever $i, j \geq 1$ and $i + j - 1 \leq n$.

Theorem B.4 *(Ky Fan inequalities) For $n \times n$ Hermitian matrices* $\mathbf{A}$ *and* $\mathbf{B}$*,*

$$\begin{aligned}&\lambda_1(\mathbf{A}+\mathbf{B}) + \cdots + \lambda_k(\mathbf{A}+\mathbf{B}) \leq \\ &\quad \lambda_1(\mathbf{A}) + \cdots + \lambda_k(\mathbf{A}) + \lambda_1(\mathbf{B}) + \cdots + \lambda_k(\mathbf{B}).\end{aligned} \tag{B.7}$$

The following Theorems B.5-B.10 are from Bai and Silverstein (2010).

Theorem B.5 *Let* $\mathbf{A}$ *and* $\mathbf{C}$ *be two $p \times n$ complex matrices. Then, for any nonnegative integers i and j, we have*

$$s_{i+j+1}(\boldsymbol{A}+\boldsymbol{C}) \leq s_{i+1}(\boldsymbol{A}) + s_{j+1}(\boldsymbol{C}) \tag{B.8}$$

Theorem B.6 *Let* $\mathbf{A}$ *and* $\mathbf{C}$ *be complex matrices of order $p \times n$ and $n \times m$. We have*

$$s_1(\mathbf{AC}) \leq s_1(\mathbf{A}) s_1(\mathbf{C}). \tag{B.9}$$

Theorem B.7 *Let* $\mathbf{A}$ *and* $\mathbf{C}$ *be complex matrices of order $p \times n$ and $n \times m$. For any $i, j \geq 0$, we have*

$$s_{i+j+1}(\mathbf{AC}) \leq s_{i+1}(\mathbf{A}) s_{j+1}(\mathbf{C}), \tag{B.10}$$

where when i>*rank*(*A*)*,* define $s_i(\mathbf{A}) = 0$.

Theorem B.8 *Let* $\mathbf{A} = (a_{ij})$ *be a complex matrix of order n and f be an increasing convex function. Then we have*

$$\sum_{j=1}^{n} f(|a_{jj}|) \leq \sum_{j=1}^{n} f(s_j(\mathbf{A})). \tag{B.11}$$

Note that when $\mathbf{A}$ is Hermitian, $s_j(\mathbf{A})$ can be replaced by eigenvalues and f need not be increasing.

Theorem B.9 *Let* $\mathbf{A}$ *and* $\mathbf{C}$ *be complex matrices of order $p \times n$ and $n \times m$. We have*

$$\sum_{j=1}^{k} s_j(\mathbf{AC}) \leq \sum_{j=1}^{k} s_j(\mathbf{A}) s_j(\mathbf{C}). \tag{B.12}$$

Theorem B.10 *Let* $\mathbf{A}$ *be a* $p \times n$ *complex matrix and* $\mathbf{U}$ *be an* $n \times m$ *complex matrix with* $\mathbf{U}^*\mathbf{U} = \mathbf{I}_m$*. Then, for any* $k \leq p$*,*

$$\sum_{j=1}^{k} s_j(\mathbf{AU}) \leq \sum_{j=1}^{k} s_j(\mathbf{A}). \tag{B.13}$$

Bibliography

Anderson, G. W., Guionnet, A., and Zeitouni, O. 2010. *An introduction to random matrices*. Cambridge Studies in Advanced Mathematics, vol. 118. Cambridge University Press, Cambridge.

Anderson, T. W., and Amemiya, Y. 1988. The asymptotic normal distribution of estimators in factor analysis under general conditions. *Ann. Statist.*, **16**(2), 759–771.

Anderson, T. W., and Rubin, H. 1956. Statistical inference in factor analysis. Pages 111–150 of: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. V.* Berkeley and Los Angeles: University of California Press.

Anderson, T.W. 2003. *An Introduction to Multivariate Statistical Analysis*. 3rd edn. Hoboken, New Jersey: John Wiley & Sons.

Arnold, L. 1967. On the asymptotic distribution of the eigenvalues of random matrices. *J. Math. Anal. Appl.*, **20**, 262–268.

Arnold, L. 1971. On Wigner's semicircle law for the eigenvalues of random matrices. *Z. Wahrsch. Verw. Gebiete*, **19**, 191–198.

Bai, Z. 1985. A note on limiting distribution of the eigenvalues of a class of random matrice. *J. of Math. Res. & Exposition*, **5**(2), 113–118.

Bai, Z. 1999. Methodologies in spectral analysis of large dimensional random matrices. A review. *Statistica Sinica*, **9**, 611–677. With comments by G. J. Rodgers and Jack W. Silverstein; and a rejoinder by the author.

Bai, Z. 2005. High dimensional data analysis. *Cosmos*, **1**(1), 17–27.

Bai, Z., and Ding, X. 2012. Estimation of spiked eigenvalues in spiked models. *Random Matrices Theory Appl.*, **1**(2), 1150011, 21.

Bai, Z., and Saranadasa, H. 1996. Effect of high dimension: By an example of a two sample problem. *Statistica Sinica*, **6**(2), 311–329.

Bai, Z., and Silverstein, J.W. 2004. CLT for linear spectral statistics of large-dimensional sample covariance matrices. *Ann. Probab.*, **32**, 553–605.

Bai, Z., and Silverstein, J.W. 2010. *Spectral Analysis of Large Dimensional Random Matrices*. 2nd edn. New York: Springer.

Bai, Z., and Yao, J. 2012. On sample eigenvalues in a generalized spiked population model. *Journal of Multivariate Analysis*, **106**, 167–177.

Bai, Z., and Yao, J.-F. 2008. Central limit theorems for eigenvalues in a spiked population model. *Annales de l'Institut Henri Poincare (B) Probabilité et Statistique*, **44**(3), 447–474.

Bai, Z., and Yin, Y. Q. 1988. A convergence to the semicircle law. *Ann. Probab.*, **16**(2), 863–875.

Bai, Z., Yin, Y. Q., and Krishnaiah, P. R. 1986. On limiting spectral distribution of product of two random matrices when the underlying distribution is isotropic. *J. Multvariate Anal.*, **19**, 189–200.

Bai, Z., Yin, Y.Q., and P.R., Krishnaiah. 1987. On the limiting empirical distribution function of the eigenvalues of a multivariate F-matrix. *The Probability Theory and Its Applications*, **32**, 490–500.

Bai, Z., Jiang, D., Yao, J.-F., and Zheng, S. 2009. Corrections to LRT on large-dimensional covariance matrix by RMT. *The Annals of Statistics*, **37**(6 B), 3822–3840.

Baik, J., and Silverstein, J.W. 2006. Eigenvalues of Large Sample Covariance Matrices of Spiked Population Models. *J. Multivariate. Anal.*, **97**, 1382–1408.

Baik, J., Ben Arous, G., and Pch, S. 2005. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.*, **33**(5), 1643–1697.

Bartlett, M. S. 1937. Properties of sufficiency arid statistical tests. *Proceedings of the Royal Society of London A*, **160**, 268–282.

Benaych-Georges, F., Guionnet, A., and Maida, M. 2011. Fluctuations of the extreme eigenvalues of finite rank deformations of random matrices. *Electron. J. Probab.*, **16**(60), 1621–1662.

Benaych-Georges, Florent, and Nadakuditi, Raj Rao. 2011. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Adv. Math.*, **227**(1), 494–521.

Billingsley, P. 1968. *Convergence of Probability Measures*. Wiley.

Birke, M., and Dette, H. 2005. A note on testing the covariance matrix for large dimension. *Statistics and Probability Letters*, **74**, 281–289.

Buja, A., Hastie, T., and Tibshirani, R. 1995. Penalized discriminant analysis. *Ann. Statist.*, **23**, 73–102.

Capitaine, Mireille, Donati-Martin, Catherine, and Féral, Delphine. 2009. The largest eigenvalues of finite rank deformation of large Wigner matrices: convergence and nonuniversality of the fluctuations. *Ann. Probab.*, **37**(1), 1–47.

Chen, Song Xi, Zhang, Li-Xin, and Zhong, Ping-Shou. 2010. Tests for high-dimensional covariance matrices. *Journal of the American Statistical Association*, **105**, 810–819.

Dempster, A. P. 1958. A high dimensional two sample significance test. *Annals of Mathematical Statististics*, **29**, 995–1010.

Féral, D., and Péché, S. 2007. The largest eigenvalue of rank one deformation of large Wigner matrices. *Comm. Math. Phys.*, **272**(1), 185–228.

Giroux, A. 2013. *Analyse Complexe (cours et exercices corrigés)*. Tech. rept. Département de mathématiques et statistique, Université de Montréal, `http://dms.umontreal.ca/~giroux/`.

Grenander, U. 1963. *Probabilities on Algebraic Structures*. New York-London: John Wiley & Sons.

Grenander, U., and Silverstein, J. 1977. Spectral analysis of networks with random topologies. *SIAM J. Appl. Math.*, **32**, 499–519.

Hastie, T., Tibshirani, R., and Friedman, J. 2009. *The Elements of Statistical Learning (2nd edition)*. Springer-Verlag.

Huber, P. J. 1973. The 1972 Wald Memorial Lecture. Robust regression: asymptotics, conjectures and Monte Carlo. *Ann. Statist.*, **35**, 73–101.

Jing, B. Y., Pan, G.M., Shao, Q.-M., and W., Zhou. 2010. Nonparametric estimate of spectral density functions of sample covariance matrices: A first step. *Ann. Statist.*, **38**, 3724–3750.

John, S. 1971. Some optimal multivariate tests. *Biometrika*, **58**, 123–127.

John, S. 1972. The distribution of a statistic used for testing sphericity of normal distributions. *Biometrika*, **59**, 169–173.

Johnstone, I. 2001. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, **29**(2), 295–327.

Johnstone, I. 2007. High dimensional statistical inference and random matrices. Pages 307–333 of: *International Congress of Mathematicians. Vol. I*. Eur. Math. Soc., Zürich.

Johnstone, I., and Titterington, D. 2009. Statistical challenges of high-dimensional data. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, **367**(1906), 4237–4253.

Jonsson, Dag. 1982. Some limit theorems for the eigenvalues of a sample covariance matrix. *J. Multivariate Anal.*, **12**(1), 1–38.

Kreĭn, M. G., and Nudel′man, A. A. 1977. *The Markov moment problem and extremal problems*. Providence, R.I.: American Mathematical Society. Ideas and problems of P. L. Čebyšev and A. A. Markov and their further development, Translated from the Russian by D. Louvish, Translations of Mathematical Monographs, Vol. 50.

Kritchman, S.and Nadler, B. 2008. Determining the number of components in a factor model from limited noisy data. *Chem. Int. Lab. Syst.*, **94**, 19–32.

Kritchman, S., and Nadler, B. 2009. Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory. *IEEE Trans. Signal Process.*, **57**(10), 3930–3941.

Ledoit, O., and Wolf, M. 2002. Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Annals of Statistics*, **30**, 1081–1102.

Lytova, A., and Pastur, L. 2009. Central limit theorem for linear eigenvalue statistics of the Wigner and the sample covariance random matrices. *Ann. Probab.*, **37**, 1778–1840.

Marčenko, V.A., and Pastur, L.A. 1967. Distribution of eigenvalues for some sets of random matrices. *Math. USSR-Sb*, **1**, 457–483.

Mehta, M.L. 2004. *Random Matrices*. third edition edn. New York: Academic Press.

Nadler, B. 2010. Nonparametric detection of signals by information theoretic criteria: performance analysis and an improved estimator. *IEEE Trans. Signal Process.*, **58**(5), 2746–2756.

Nagao, H. 1973a. Asymptotic expansions of the distributions of Bartlett's test and sphericity test under the local alternatives. *Annals of the Institute of Statistical Mathematics*, **25**, 407–422.

Nagao, H. 1973b. On some test criteria for covariance matrix. *Annals of Statistics*, **1**, 700–709.

Nica, A., and Speicher, R. 2006. *Lectures on the Combinatorics of Free Probability*. New York: Cambridge University Press.

Pan, G. 2014. Comparison between two types of large sample covariance matrices. *Annales de l'institut Henri Poincare (B) Probability and Statistics*, **50**(2), 655–677.

Pan, G.M., and Zhou, W. 2008. Central limit theorem for signal-to-interference ratio of reduced rank linear receiver. *Ann. Appl. Probab.*, 1232–1270.

Passemier, D., and Yao, J. 2012. On determining the number of spikes in a high-dimensional spiked population model. *Random Matrix: Theory and Applciations*, **1**, 1150002.

Passemier, D., and Yao, J. 2014. Estimation of the number of spikes, possibly equal, in the high-dimensional case. *Journal of Multivariate Analysis*, **127**, 173–183.

Passemier, D., Li, Z., and Yao, J. 2017. On estimation of the noise variance in high-dimensional probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B: Methodological*, **79**(1), 51–67.

Pastur, L., and Shcherbina, M. 2011. *Eigenvalue distribution of large random matrices*. Mathematical Surveys and Monographs, vol. 171. American Mathematical Society, Providence, RI.

Pastur, L. A. 1972. On the spectrum of random matrices. *Teoret. Mat. Phys.*, **10**, 67–74.

Pastur, L. A. 1973. Spectra of random self-adjoint operators. *Russian Math. Surv.*, **28**, 1–67.

Paul, Debashis. 2007. Asymptotics of sample eigenstruture for a large dimensional spiked covariance mode. *Statistica Sinica*, **17**, 1617–1642.

Péché, S. 2006. The largest eigenvalue of small rank perturbations of Hermitian random matrices. *Probab. Theory Related Fields*, **134**(1), 127–173.

Pizzo, A., Renfrew, D., and Soshnikov, A. 2013. On finite rank deformations of Wigner matrices. *Ann. Inst. Henri Poincaré Probab. Stat.*, **49**(1), 64–94.

Renfrew, D., and Soshnikov, A. 2013. On finite rank deformations of Wigner matrices II: Delocalized perturbations. *Random Matrices Theory Appl.*, **2**(1), 1250015, 36.

Sheather, S. J., and Jones, M. C. 1991. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society series B*, **53**, 683–690.

Silverstein, J. W., and Combettes, P. L. 1992. Signal detection via spectral theory of large dimensional random matrices. *IEEE Trans. Signal Processing*, **40**, 2100–2104.

Silverstein, Jack W. 1985. The limiting eigenvalue distribution of a multivariate *F* matrix. *SIAM J. Math. Anal.*, **16**(3), 641–646.

Silverstein, Jack W., and Choi, Sang-Il. 1995. Analysis of the limiting spectral distribution of large-dimensional random matrices. *J. Multivariate Anal.*, **54**(2), 295–309.

Srivastava, M.S., Kollo, T., and von Rosen, D. 2011. Some tests for the covariance matrix with fewer observations than the dimension under non-normality. *J. Multivariate. Anal.*, **102**, 1090–1103.

Srivastava, Muni S. 2005. Some tests concerning the covariance matrix in high dimensional data. *Journal of Japan Statistical Society*, **35**(2), 251–272.

Sugiura, N., and Nagao, H. 1968. Unbiasedness of some test criteria for the equality of one or two covariance matrices. *Annals of Mathematical Statistics*, **39**, 1686–1692.

Tao, T. 2012. *Topics in random matrix theory*. Graduate Studies in Mathematics, vol. 132. American Mathematical Society, Providence, RI.

Ulfarsson, M. O., and Solo, V. 2008. Dimension estimation in noisy PCA with SURE and random matrix theory. *IEEE Trans. Signal Process.*, **56**(12), 5804–5816.

Wachter, Kenneth W. 1978. The strong limits of random matrix spectra for sample matrices of independent elements. *Ann. Probability*, **6**(1), 1–18.

Wachter, Kenneth W. 1980. The limiting empirical measure of multiple discriminant ratios. *Ann. Statist.*, **8**, 937–957.

Wang, Q., and Yao, J. 2013. On the sphericity test with large-dimensional observations. *Electronic Journal of Statistics*, **7**(1), 2164–2192.

Wax, M., and Kailath, Th. 1985. Detection of signals by information theoretic criteria. *IEEE Trans. Acoust. Speech Signal Process.*, **33**(2), 387–392.

Wigner, E.P. 1955. Characteristic vectors bordered matrices with infinite dimensions. *The Annals of Mathematics*, **62**, 548–564.

Wigner, E.P. 1958. On the distributions of the roots of certain symmetric matrices. *The Annals of Mathematics*, **67**, 325–327.

Yin, Y. Q. 1986. Limiting spectral distribution for a class of random matrices. *J. Multivariate Anal.*, **20**(1), 50–68.

Yin, Y. Q., and Krishnaiah, P. R. 1983. A limit theorem for the eigenvalues of product of two random matrices. *J. Multivariate Anal.*, **13**, 489–507.

Zheng, S. 2012. Central Limit Theorem for Linear Spectral Statistics of Large Dimensional *F* Matrix. *Ann. Institut Henri Poincaré Probab. Statist.*, **48**, 444–476.

Zheng, S., Jiang, D., Bai, Z., and He, X. 2014. Inference on multiple correlation coefficients with moderately high dimensional data. *Biometrika*, **101**(3), 748–754.

Zheng, S., Bai, Z., and Yao, J. 2015. Substitution principle for CLT of linear spectral statistics of high-dimensional sample covariance matrices with applications to hypothesis testing. *The Annals of Statistics*, **43**(2), 546–591.

Index

Marčenko-Pastur distribution, 10
Marčenko-Pastur equation, 18
multiple correlation coefficient, 52

adjusted multiple correlation coefficient, 53
adjusted sample size, 44, 45
analytic function, 129

bulk of the spectrum, 86

Cauchy interlacing law, 136
Cauchy-Riemann theorem, 129
CLT for spectral statistics
  Bai and Silverstein's CLT, 39
  of covariance matrix, 31
  of Fisher matrix, 41
contour integral
  method of residue, 135
curve, 130
  anticlockwise run, 130
  clockwise run, 130
  closed curve, 130
  contour, 130
  differentiable curve, 130
  exterior of a closed curve, 130
  interior of a closed curve, 130
curvilinear integral, 130

detection of the number of signals, 112
domain, 130
  simply connected, 132

elbow rule in a scree plot, 86
empirical spectral distribution, 8

Fisher LSD, 23
Fisher matrix, 7, 23
fixed-point algorithm for LSD, 21
fundamental spikes, 91

generalised Marčenko-Pastur distributions, 18
generalised spike, 89
generalised spiked population model, 89, 128
generalised variance, 7, 30, 47

holomorphic function, 129
  Cauchy theorem, 132
  essential singularity, 133
  Laurent series, 132
  pole, 133
  removable singularity, 133
  residue, 134

Johnstone's spiked population model, 88, 96, 113, 127

Kritchman and Nadler's detector, 115
Ky Fan inequalities, 137

limiting spectral distribution, 8
linear spectral statistic, 7, 30

meromorphic function, 133
multiple correlation coefficient, 52

non-fundamental spikes, 91
non-null case, 88
null case, 87

phoneme data set, 127
population spectral distribution
  definition, 18
  discrete, 18
  moment map, 19

random sesquilinear form, 97

sample multiple correlation coefficient, 52
sample covariance matrix, 7
  centerd, 44
  non-centerd, 44
signal strengths, 112
Silverstein equation, 18
singular value decomposition, 136
spectral decomposition, 136
spike eigenvalues, 88
spiked eigenvectors, 94
spiked population model, 88
spiked sample eigenvalues, 91
Standard and Poor 500 data, 86
Stieltjes transform, 9
  inversion formula, 9
substitution principle, 45
  application example, 68, 77
  for covariance matrix, 45
  for Fisher matrix, 46

testing on covariance matrix
  Chen's test for sphericity, 80
  corrected John's test for sphericity, 80
  corrected likelihood ratio for equality, 66, 68

corrected likelihood ratio for sphericity, 77
equality test, 58
identity test, 64
John's test for sphericity, 64, 75
large-dimensional equality test, 65, 67, 69–71
large-dimensional sphericity test, 74
likelihood ratio for sphericity, 63, 75
Nagao's test for sphericity, 64, 75
sphericity test, 62
Tracy-Widom law, 88, 119

Weyl inequalities, 137
Wishart distribution, 48