# Spectrum Estimation: A Unified Framework for Covariance Matrix Estimation and PCA in Large Dimensions*

Olivier Ledoit
Department of Economics
University of Zurich
CH-8032 Zurich, Switzerland
olivier.ledoit@econ.uzh.ch

Michael Wolf†
Department of Economics
University of Zurich
CH-8032 Zurich, Switzerland
michael.wolf@econ.uzh.ch

First version: January 2013
This version: July 2013

**Abstract**

Covariance matrix estimation and principal component analysis (PCA) are two cornerstones of multivariate analysis. Classic textbook solutions perform poorly when the dimension of the data is of a magnitude similar to the sample size, or even larger. In such settings, there is a common remedy for both statistical problems: *nonlinear shrinkage* of the eigenvalues of the sample covariance matrix. The optimal nonlinear shrinkage formula depends on unknown population quantities and is thus not available. It is, however, possible to consistently estimate an oracle nonlinear shrinkage, which is motivated on asymptotic grounds. A key tool to this end is consistent estimation of the set of eigenvalues of the population covariance matrix (also known as the *spectrum*), an interesting and challenging problem in its own right. Extensive Monte Carlo simulations demonstrate that our methods have desirable finite-sample properties and outperform previous proposals.

KEY WORDS: Large-dimensional asymptotics, covariance matrix eigenvalues, nonlinear shrinkage, principal component analysis.

JEL CLASSIFICATION NOS: C13.

*This work was partially completed while the authors were visiting the Institute for Mathematical Sciences, National University of Singapore in 2012. The visit was supported by the Institute.

†Research has been supported by the NCCR Finrisk project "New Methods in Theoretical and Empirical Asset Pricing".

# 1 Introduction

This paper tackles three important problems in multivariate statistics: 1) the estimation of the eigenvalues of the covariance matrix; 2) the estimation of the covariance matrix itself; and 3) principal component analysis (PCA). In many modern applications, the matrix dimension is not negligible with respect to the sample size, so textbook solutions based on classic (fixed-dimension) asymptotics are no longer appropriate. A better-suited framework is *large-dimensional asymptotics*, where the matrix dimension and the sample size go to infinity together, while their ratio — called the *concentration* — converges to a finite, nonzero limit. Under large-dimensional asymptotics, the sample covariance matrix is no longer consistent, and neither are its eigenvalues nor its eigenvectors.

One of the interesting features of large-dimensional asymptotics is that principal component analysis can no longer be conducted using covariance matrix eigenvalues. The variation explained by a principal component is not equal to the corresponding sample eigenvalue and — perhaps more surprisingly — it is not equal to the corresponding population eigenvalue either. To the best of our knowledge, this fact has not been noticed before. The variation explained by a principal component is obtained instead by applying a *nonlinear shrinkage* formula to the corresponding sample eigenvalue. This nonlinear shrinkage formula depends on the unobservable population covariance matrix, but thankfully it can approximated by an *oracle shrinkage* formula which depends 'only' on the unobservable eigenvalues of the population covariance matrix. This is the connection with the first of the three problems mentioned above. Once we have a consistent estimator of the population eigenvalues, we can use it to derive a consistent estimator of the oracle shrinkage.

The connection with the second problem, the estimation of the whole covariance matrix, is that the nonlinear shrinkage formula that gives the variation explained by a principal component also yields the optimal rotation-equivariant estimator of the covariance matrix according to the Frobenius norm. Thus, if we can consistently estimate the population eigenvalues, and if we plug them into the oracle shrinkage formula, we can address the problems of PCA and covariance matrix estimation in a unified framework.

It needs to be pointed out here that in a rotation-equivariant framework, consistent (or even improved) estimators of the population eigenvectors are not available; instead, one needs to retain the sample eigenvectors. As a consequence, consistent estimation of the population covariance matrix itself is not possible. Nevertheless, a rotation-equivariant estimator can still be very useful for practical applications, as evidenced by the popularity of the previous proposal of Ledoit and Wolf (2004). An alternative approach, which allows for consistent estimation of the population covariance matrix under suitable regularity conditions, is to impose additional structure on the estimation problem, such as sparseness, a graph model, or an (approximate) factor model. But whether such structure does indeed exist is something that cannot be verified from the data. Therefore, at least in some applications, a structure-free approach will be preferred by applied researchers. This is the problem that we address, aiming to further improve upon Ledoit and Wolf (2004).

Of course estimating population eigenvalues consistently under large-dimensional asymptotics is no trivial matter. Until recently, most researchers in the field even feared it might be impossible because deducing population eigenvalues from sample eigenvalues showed some symptoms of *ill-posedness*. This means that small estimation errors in the sample eigenvalues would be amplified by the specific mathematical structure of the asymptotic relationship between sample and population eigenvalues. But two recent articles by Mestre (2008) and El Karoui (2008) challenged this widely-held belief and gave some hope that it might be possible after all to estimate the population eigenvalues consistently. Still, a general satisfactory solution is not available to date.

The work of Mestre (2008) only applies when the number of distinct population eigenvalues remains finite as matrix dimension goes to infinity. In practice, this means that the number of distinct eigenvalues must be negligible with respect to the total number of eigenvalues. As a further restriction, the number of distinct eigenvalues and their multiplicities must be known. The only unknown quantities to be estimated are the locations of the eigenvalues; of course, this is still a difficult task. Yao et al. (2012) propose a more general estimation procedure that does not require knowledge of the multiplicities, though it still requires knowledge of the number of distinct population eigenvalues. This setting is too restrictive for many applications.

The method developed by El Karoui (2008) allows for an arbitrary set of population eigenvalues, but does not appear to have good finite-sample properties. In fact, our simulations seem to indicate that this estimator is not even consistent; see Section 5.1.1.

The first contribution of the present paper is, therefore, to develop an estimator of the population eigenvalues that is consistent under large-dimensional asymptotics *regardless* of whether or not they are clustered, and that also performs well in finite sample. This is achieved through a more precise characterization of the asymptotic behavior of sample eigenvalues. Whereas existing results only specify how the eigenvalues behave on average, namely, how many fall in any given interval, we determine *individual* limits.

Our second contribution is to show how this consistent estimator of population eigenvalues can be used for improved estimation of the covariance matrix when the dimension is large compared to the sample size. This was already considered in Ledoit and Wolf (2012), but only in the limited setup where the dimension is smaller than the sample size. Thanks to the advances introduced in the present paper, we can also handle the more difficult case where the dimension exceeds the sample size and the sample covariance matrix is singular.

Our third and final contribution is to show how the same nonlinear shrinkage formula can be used to estimate the fraction of variation explained by a given collection of principal components in PCA, which is key in deciding how many principal components to retain.

The remainder of the paper is organized as follows. Section 2 presents our estimator of the eigenvalues of the population covariance matrix under large-dimensional asymptotics. Section 3 discusses covariance matrix estimation, and Section 4 principal component analysis. Section 5 studies finite-sample performance via Monte Carlo simulations. Section 6 provides a brief empirical application of PCA to stock return data. Section 7 concludes. The proofs of all mathematical results are collected in the appendix.

# 2 Estimation of Population Covariance Matrix Eigenvalues

## 2.1 Large-Dimensional Asymptotics and Basic Framework

Let $n$ denote the sample size and $p := p(n)$ the number of variables. It is assumed that the ratio $p/n$ converges as $n \to \infty$ to a limit $c \in (0,1) \cup (1,\infty)$ called the *concentration*. The case $c = 1$ is ruled out for technical reasons. We make the following assumptions.

(A1) The population covariance matrix $\Sigma_n$ is a nonrandom $p$-dimensional positive definite matrix.

(A2) $X_n$ is an $n \times p$ matrix of real independent and identically distributed (i.i.d.) random variables with zero mean, unit variance, and finite fourth moment. One only observes $Y_n := X_n \Sigma_n^{1/2}$, so neither $X_n$ nor $\Sigma_n$ are observed on their own.

(A3) $\boldsymbol{\tau}_n := (\tau_{n,1}, \ldots, \tau_{n,p})'$ denotes a system of eigenvalues of $\Sigma_n$, sorted in increasing order, and $(v_{n,1}, \ldots, v_{n,p})$ denotes an associated system of eigenvectors. The empirical distribution function (e.d.f.) of the population eigenvalues is defined as: $\forall t \in \mathbb{R}$, $H_n(t) := p^{-1} \sum_{i=1}^{p} \mathbb{1}_{[\tau_{n,i}, +\infty)}(t)$, where $\mathbb{1}$ denotes the indicator function of a set. $H_n$ is called the *spectral distribution (function)*. It is assumed that $H_n$ converges weakly to a limit law $H$, called the *limiting spectral distribution (function)*.

(A4) $\mathsf{Supp}(H)$, the support of $H$, is the union of a finite number of closed intervals, bounded away from zero and infinity. Furthermore, there exists a compact interval in $(0,\infty)$ that contains $\mathsf{Supp}(H_n)$ for all $n$ large enough.

Let $\boldsymbol{\lambda}_n := (\lambda_{n,1}, \ldots, \lambda_{n,p})'$ denote a system of eigenvalues of the sample covariance matrix $S_n := n^{-1} Y_n' Y_n = n^{-1} \Sigma_n^{1/2} X_n' X_n \Sigma_n^{1/2}$, sorted in increasing order, and let $(u_{n,1}, \ldots, u_{n,p})$ denote an associated system of eigenvectors. The first subscript, $n$, may be omitted when no confusion is possible. The e.d.f. of the sample eigenvalues is defined as: $\forall t \in \mathbb{R}$, $F_n(t) := p^{-1} \sum_{i=1}^{p} \mathbb{1}_{[\lambda_i, +\infty)}(t)$. The literature on the eigenvalues of sample covariance matrices under large-dimensional asymptotics — also known as *random matrix theory* (RMT) literature — is based on a foundational result due to Marčenko and Pastur (1967). It has been strengthened and broadened by subsequent authors including Silverstein (1995), Silverstein and Bai (1995), Silverstein and Choi (1995), and Bai and Silverstein (1998, 1999), among others. These articles imply that there exists a limiting sample spectral distribution $F$ such that

$$\forall x \in \mathbb{R} \setminus \{0\} \qquad F_n(x) \xrightarrow{\text{a.s.}} F(x) \ . \tag{2.1}$$

In other words, the *average* number of sample eigenvalues falling in any given interval is known asymptotically.

In addition, the existing literature has unearthed important information about the limiting distribution $F$. Silverstein and Choi (1995) show that $F$ is everywhere continuous except (potentially) at zero, and that the mass that $F$ places at zero is given by

$$F(0) = \max\Big\{1 - \frac{1}{c}, H(0)\Big\} \ . \tag{2.2}$$

Furthermore, there is a seminal equation relating $F$ to $H$ and $c$. Some additional notation is required to present this equation.

For any nondecreasing function $G$ on the real line, $m_G$ denotes the *Stieltjes transform* of $G$:

$$\forall z \in \mathbb{C}^+ \qquad m_G(z) := \int \frac{1}{\lambda - z}\, dG(\lambda) \ ,$$

where $\mathbb{C}^+$ denotes the half-plane of complex numbers with strictly positive imaginary part.

The Stieltjes transform admits a well-known inversion formula:

$$G(b) - G(a) = \lim_{\eta \to 0^+} \frac{1}{\pi} \int_a^b \mathsf{Im}\big[m_G(\xi + i\eta)\big] d\xi \ , \tag{2.3}$$

if $G$ is continuous at $a$ and $b$. Here, and in the remainder of the paper, we shall use the notations $\mathsf{Re}(z)$ and $\mathsf{Im}(z)$ for the real and imaginary parts, respectively, of a complex number $z$, so that

$$\forall z \in \mathbb{C} \qquad z = \mathsf{Re}(z) + i \cdot \mathsf{Im}(z) \ .$$

The most elegant version of the equation relating $F$ to $H$ and $c$, due to Silverstein (1995), states that $m := m_F(z)$ is the unique solution in the set

$$\left\{ m \in \mathbb{C} : -\frac{1-c}{z} + cm \in \mathbb{C}^+ \right\} \tag{2.4}$$

to the equation

$$\forall z \in \mathbb{C}^+ \qquad m_F(z) = \int \frac{1}{\tau\big[1 - c - c\, z\, m_F(z)\big] - z}\, dH(\tau) \ . \tag{2.5}$$

As explained, the Stieltjes transform of $F$, $m_F$, is a function whose domain is the upper half of the complex plane. It can be extended to the real line, since Silverstein and Choi (1995) show that: $\forall \lambda \in \mathbb{R} \setminus \{0\}$, $\lim_{z \in \mathbb{C}^+ \to \lambda} m_F(z) =: \breve{m}_F(\lambda)$ exists. When $c < 1$, $\breve{m}_F(0)$ also exists and $F$ has a continuous derivative $F' = \pi^{-1} \mathsf{Im}\left[\breve{m}_F\right]$ on all of $\mathbb{R}$ with $F' \equiv 0$ on $(-\infty, 0]$. (One should remember that although the argument of $\breve{m}_F$ is real-valued now, the output of the function is still a complex number.)

For purposes that will become apparent later, it is useful to reformulate equation (2.5). The limiting e.d.f. of the eigenvalues of $n^{-1} Y_n' Y_n = n^{-1} \Sigma_n^{1/2} X_n' X_n \Sigma_n^{1/2}$ was defined as $F$. In addition, define the limiting e.d.f. of the eigenvalues of $n^{-1} Y_n Y_n' = n^{-1} X_n \Sigma_n X_n'$ as $\underline{F}$; note that the eigenvalues of $n^{-1} Y_n' Y_n$ and $n^{-1} Y_n Y_n'$ only differ by $|n - p|$ zero eigenvalues. It then holds:

$$\forall x \in \mathbb{R} \qquad \underline{F}(x) = (1 - c)\, \mathbb{1}_{[0,\infty)}(x) + c\, F(x) \tag{2.6}$$

$$\forall x \in \mathbb{R} \qquad F(x) = \frac{c-1}{c} \mathbb{1}_{[0,\infty)}(x) + \frac{1}{c}\, \underline{F}(x) \tag{2.7}$$

$$\forall z \in \mathbb{C}^+ \qquad m_{\underline{F}}(z) = \frac{c-1}{z} + c\, m_F(z) \tag{2.8}$$

$$\forall z \in \mathbb{C}^+ \qquad m_F(z) = \frac{1-c}{c\, z} + \frac{1}{c}\, m_{\underline{F}}(z) \ . \tag{2.9}$$

(Recall here that $F$ has mass $(c-1)/c$ at zero when $c > 1$, so that both $F$ and $\underline{F}$ are nonnegative functions indeed for any value $c > 0$.)

With this notation, equation (1.13) of Marčenko and Pastur (1967) reframes equation (2.5) as: for each $z \in \mathbb{C}^+$, $m := m_{\underline{F}}(z)$ is the unique solution in $\mathbb{C}^+$ to the equation

$$m = -\left[z - c \int \frac{\tau}{1+\tau\, m} dH(\tau)\right]^{-1} . \tag{2.10}$$

While in the case $c < 1$, $\breve{m}_F(0)$ exists and $F$ is continuously differentiable on all of $\mathbb{R}$, as mentioned above, in the case $c > 1$, $\breve{m}_{\underline{F}}(0)$ exists and $\underline{F}$ is continuously differentiable on all of $\mathbb{R}$.

## 2.2 Individual Behavior of Sample Eigenvalues: the QuEST Function

We introduce a nonrandom multivariate function called the *Quantized Eigenvalues Sampling Transform*, or QuEST for short, which discretizes, or *quantizes*, the relationship between $F$, $H$, and $c$ defined in equations (2.1)–(2.3). For any positive integers $n$ and $p$, the QuEST function, denoted by $Q_{n,p}$, is defined as

$$Q_{n,p} : [0,\infty)^p \longrightarrow [0,\infty)^p \tag{2.11}$$

$$\mathbf{t} := (t_1, \ldots, t_p)' \longmapsto Q_{n,p}(\mathbf{t}) := \left(q^1_{n,p}(\mathbf{t}), \ldots, q^p_{n,p}(\mathbf{t})\right)' , \tag{2.12}$$

where $\forall z \in \mathbb{C}^+ \quad m := m^{\mathbf{t}}_{n,p}(z)$ is the unique solution in the set

$$\left\{ m \in \mathbb{C} : -\frac{n-p}{nz} + \frac{p}{n}\, m \in \mathbb{C}^+ \right\} \tag{2.13}$$

to the equation

$$m = \frac{1}{p} \sum_{i=1}^{p} \frac{1}{t_i \left(1 - \frac{p}{n} - \frac{p}{n}\, z\, m\right) - z} , \tag{2.14}$$

$$\forall x \in \mathbb{R} \qquad F^{\mathbf{t}}_{n,p}(x) := \begin{cases} \max\left\{1 - \frac{n}{p}, \frac{1}{p}\sum_{i=1}^{p} \mathbb{1}_{\{t_i=0\}}\right\} & \text{if } x = 0 , \\ \lim_{\eta \to 0^+} \frac{1}{\pi} \int_{-\infty}^{x} \mathsf{Im}\left[m^{\mathbf{t}}_{n,p}(\xi + i\eta)\right]\, d\xi & \text{otherwise} , \end{cases} \tag{2.15}$$

$$\forall u \in [0,1] \qquad \left(F^{\mathbf{t}}_{n,p}\right)^{-1}(u) := \sup\{x \in \mathbb{R} : F^{\mathbf{t}}_{n,p}(x) \le u\} , \tag{2.16}$$

$$\text{and} \qquad \forall i = 1, \ldots, p \qquad q^i_{n,p}(\mathbf{t}) := p \int_{(i-1)/p}^{i/p} \left(F^{\mathbf{t}}_{n,p}\right)^{-1}(u)\, du . \tag{2.17}$$

It is obvious that equation (2.13) quantizes equation (2.4), that equation (2.14) quantizes equation (2.5), and that equation (2.15) quantizes equations (2.2) and (2.3). Thus, $F^{\mathbf{t}}_{n,p}$ is the limiting distribution (function) of sample eigenvalues corresponding to the population spectral distribution (function) $p^{-1}\sum_{i=1}^{p} \mathbb{1}_{[t_i,+\infty)}$. Furthermore, by equation (2.16), $\left(F^{\mathbf{t}}_{n,p}\right)^{-1}$ represents the inverse spectral distribution function, also known as the *quantile* function.

**Remark 2.1** (Definition of Quantiles)**.** The standard definition of the $(i-0.5)/p$ quantile of $F_{n,p}^{\mathbf{t}}$ is $\left(F_{n,p}^{\mathbf{t}}\right)^{-1}((i-0.5)/p)$, where $\left(F_{n,p}^{\mathbf{t}}\right)^{-1}$ is defined in equation (2.16). It turns out, however, that the 'smoothed' version $q_{n,p}^{i}(\mathbf{t})$ given in equation (2.17) leads to improved accuracy, higher stability, and faster computations of our numerical algorithm, to be detailed below, in practice.

Since $F_n$ is an empirical distribution (function), its quantiles are not uniquely defined. For example, the statistical software R offers nine different versions of sample quantiles in its function `quantile`; version 5 corresponds to our convention of considering $\lambda_{n,i}$ as the $(i-0.5)/p$ quantile of $F_n$. ■

Consequently, a set of $(i-0.5)/p$ quantiles $(i=1,\ldots,p)$ is given by $Q_{n,p}(\mathbf{t})$ for $F_{n,p}^{\mathbf{t}}$ and is given by $\boldsymbol{\lambda}_n$ for $F_n$. The relationship between $Q_{n,p}(\mathbf{t})$ and $\boldsymbol{\lambda}_n$ is further elucidated by the following theorem.

**Theorem 2.1.** *If Assumptions (A1)–(A4) are satisfied, then*

$$\frac{1}{p}\sum_{i=1}^{p}\left[q_{n,p}^{i}(\boldsymbol{\tau}_n)-\lambda_{n,i}\right]^2 \xrightarrow{\text{a.s.}} 0\ . \tag{2.18}$$

Theorem 2.1 states that the sample eigenvalues converge individually to their nonrandom QuEST function counterparts. This individual notion of convergence is defined as the Euclidian distance between the vectors $\boldsymbol{\lambda}_n$ and $Q_{n,p}(\boldsymbol{\tau}_n)$, normalized by the matrix dimension $p$. It is the appropriate normalization because, as $p$ goes to infinity, the left-hand side of equation (2.18) approximates the $L^2$ distance between the functions $F_n^{-1}$ and $\left(F_{n,p}^{\boldsymbol{\tau}_n}\right)^{-1}$. This metric can be thought of as a 'cross-sectional' mean squared error, in the same way that $F_n$ is a cross-sectional distribution function.

Theorem 2.1 improves over the well-known results from the random matrix theory literature reviewed in Section 2.1 in two significant ways.

1) It is based on the $p$ population eigenvalues $\boldsymbol{\tau}_n$, not the limiting spectral distribution $H$. Dealing with $\boldsymbol{\tau}_n$ (or, equivalently, $H_n$) is straightforward because it is integral to the actual data-generating process; whereas dealing with $H$ is more delicate because we do not know how $H_n$ converges to $H$. Also there are potentially different $H$'s that $H_n$ could converge to, depending on what we assume will happen as the dimension increases.

2) Theorem 2.1 characterizes the *individual* behavior of the sample eigenvalues, whereas equation (2.1) only characterizes their *average* behavior, namely, what proportion falls in any given interval. Individual results are more precise than average results. Thus, Theorem 2.1 shows that the sample eigenvalues are better behaved under large-dimensional asymptotics than previously thought.

Both of these improvements are made possible thanks to the introduction of the QuEST function. In spite of the apparent complexity of the mathematical definition of the QuEST function, it can be computed quickly and efficiently along with its analytical Jacobian as evidenced by Figure 1, and it behaves well numerically.

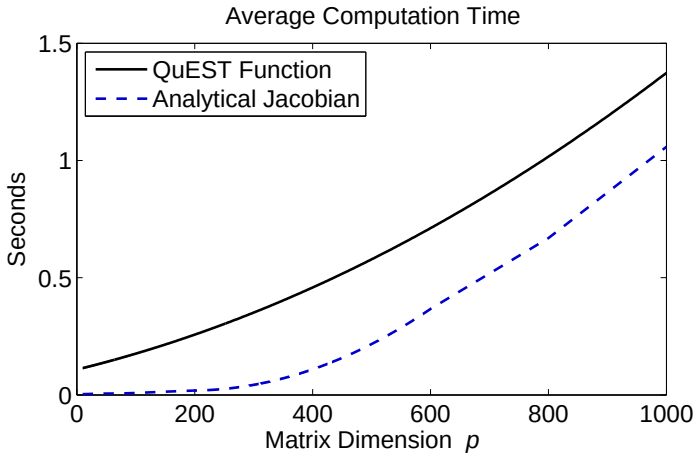



Figure 1: Average computation time for the QuEST function and its analytical Jacobian. The setup is the same as in Figure 2 below. The QuEST function and its analytical Jacobian are programmed in Matlab. The computer is a 2.4 GHz Quad-Core Intel Xeon desktop Mac.

### 2.3 Consistent Estimation of Population Eigenvalues

Once the truth of Theorem 2.1 has been established, it becomes tempting to construct an estimator of population covariance matrix eigenvalues simply by minimizing the expression on the left-hand side of equation (2.18) over all possible sets of population eigenvectors. This is exactly what we do in the following theorem.

**Theorem 2.2.** *Suppose that Assumptions (A1)–(A4) are satisfied. Define*

$$\widehat{\boldsymbol{\tau}}_n := \underset{\mathbf{t}\in[0,\infty)^p}{\operatorname{argmin}} \frac{1}{p}\sum_{i=1}^{p}\left[q_{n,p}^i(\mathbf{t}) - \lambda_{n,i}\right]^2, \tag{2.19}$$

*where* $\boldsymbol{\lambda}_n := (\lambda_{n,1},\ldots,\lambda_{n,p})'$ *are the sample covariance matrix eigenvalues, and* $Q_{n,p}(\mathbf{t}) := \left(q_{n,p}^1(\mathbf{t}),\ldots,q_{n,p}^p(\mathbf{t})\right)'$ *is the nonrandom QuEST function defined in equations* (2.11)–(2.14)*; both* $\widehat{\boldsymbol{\tau}}_n$ *and* $\boldsymbol{\lambda}_n$ *are assumed sorted in increasing order. Let* $\widehat{\tau}_{n,i}$ *denote the* $i$*th entry of* $\widehat{\boldsymbol{\tau}}_n$ $(i = 1,\ldots,p)$*, and let* $\boldsymbol{\tau}_n := (\tau_{n,1},\ldots,\tau_{n,p})'$ *denote the population covariance matrix eigenvalues sorted in increasing order. Then*

$$\frac{1}{p}\sum_{i=1}^{p}\left[\widehat{\tau}_{n,i} - \tau_{n,i}\right]^2 \xrightarrow{\text{a.s.}} 0\ . \tag{2.20}$$

Theorem 2.2 shows that the estimated eigenvalues converge *individually* to the population eigenvalues, in the same sense as above, using the dimension-normalized Euclidian distance.

**Remark 2.2.** Mathematically speaking, equation (2.19) performs two tasks: it projects $\boldsymbol{\lambda}_n$ onto the image of the QuEST function, and then inverts the QuEST function. Since the image of the QuEST function is a strict subset of $[0,\infty)^p$, $\boldsymbol{\lambda}_n$ will generally be outside of it. It is the first of these two tasks that gets around any potential ill-posedness by *regularizing* the set of observed sample eigenvalues.

Practically speaking, both tasks are performed simultaneously by a nonlinear optimizer. We use a standard off-the-shelf commercial software called SNOPT$^{\text{TM}}$ (Version 7.4), see Gill et al. (2002), but other choices may work well too. ■

## 2.4 Comparison with Other Approaches

El Karoui (2008) also attempts to discretize equation (2.5) and invert it, but he opts for a completely opposite method of discretization which does not exploit the natural discreteness of the population spectral distribution for finite $p$. If the population spectral distribution $H_n$ is approximated by a convex linear combination of step functions

$$\forall x \in \mathbb{R} \quad \widetilde{H}(x) := \sum_{i=1}^{p} w_i \mathbb{1}_{\{x \geq t_i\}} \quad \text{where} \quad \forall i = 1, \ldots, p \quad t_i \geq 0,\ w_i \geq 0 \text{ , and } \sum_{i=1}^{p} w_i = 1 \text{ ,}$$

then in the optimization problem (2.19), we keep the weights $w_i$ $(i = 1, \ldots, p)$ fixed at $1/p$ while varying the location parameters $t_i$ $(i = 1, \ldots, p)$. In contrast, El Karoui (2008) does exactly the reverse: he keeps the location parameters $t_i$ fixed on a grid while varying the weights $w_i$. Thus, El Karoui (2008) projects the population spectral distribution onto a "dictionary". Furthermore, instead of matching population eigenvalues to sample eigenvalues on $\mathbb{R}$, he matches a function of $m_{\widetilde{H}}$ to a function of $m_{F_n}$ on $\mathbb{C}^+$, which makes his algorithm relatively complicated; see Ledoit and Wolf (2012, pages 1043–1044). Despite our best efforts, we were unable to replicate his convergence results in Monte Carlo simulations: in our implementation, his estimator performs poorly overall and does not even appear to be consistent; see Section 5.1.1. Unless someone circulates an implementation of the algorithm described in El Karoui (2008) that works, we have to write off this approach as impractical.

Another related paper is the one by Ledoit and Wolf (2012). They use the same discretization strategy as El Karoui (2008) (fix location parameters and vary weights) but, as we do here, match population eigenvalues to sample eigenvalues on the real line. Unlike we do here, they measure closeness by a sup-distance rather than by the Euclidean distance. Ledoit and Wolf (2012) only consider the case $p < n$. Unfortunately, their nonlinear optimizer no longer converges reliably in the case $p > n$, as we found out in subsequent experiments; this necessitated the development of the alternative discretization strategy described above, as well as the change from sup-distance to Euclidean distance to measure closeness.

Furthermore, Ledoit and Wolf (2012) are not directly interested in estimating the population eigenvalues; it is just an intermediary step towards their ultimate objective, which is the estimation of the covariance matrix itself. Therefore they do not report any Monte Carlo simulations on the finite-sample behavior of their estimator of the population eigenvalues.

In any case, the aim of the present paper is to develop an estimator of the population eigenvalues that works also when $p > n$, so the approach of Ledoit and Wolf (2012) is ruled out. The different discretization strategy that we employ here, together with the alternative distance measure, enables us to construct an estimator of $\boldsymbol{\tau}_n$ that works across both cases $p < n$ and $p > n$. It is important to point out that the new estimator of population eigenvalues

is not only more general, in the sense that it also works for the case $p > n$, but it also works better for the case $p < n$; see Section 5.2.

As for the papers of Mestre (2008) and Yao et al. (2012), their methods are based on contour integration of analytic functions in the complex plane. They can only extract a finite number $\bar{M}$ of functionals of $H_n$, such as the locations of high-multiplicity eigenvalue clusters, or the trace of powers of $\Sigma_n$. The main difference with our method is that we extract many more items of information: namely, $p$ population eigenvalues. This distinction is crucial because the ratio $\bar{M}/p$ vanishes asymptotically. It explains why we are able to recover the whole population spectrum in the general case, whereas they are not.

# 3 Covariance Matrix Estimation

The estimation of the covariance matrix $\Sigma_n$ is already considered by Ledoit and Wolf (2012), but only for the case $p < n$. In particular, they propose a nonlinear shrinkage approach, which we will now extend to the case $p > n$. To save space, the reader is referred to their paper for a more detailed discussion of the nonlinear shrinkage methodology and a comparison to other estimation strategies of large-dimensional covariance matrices, such as the linear shrinkage estimator of Ledoit and Wolf (2004).

## 3.1 Oracle Shrinkage

The starting point is to restrict attention to *rotation-equivariant* estimators of $\Sigma_n$. To be more specific, let $W$ be an arbitrary $p$-dimensional rotation matrix. Let $\widehat{\Sigma}_n := \widehat{\Sigma}_n(Y_n)$ be an estimator of $\Sigma_n$. Then the estimator is said to be *rotation-equivariant* if it satisfies $\widehat{\Sigma}_n(Y_nW) = W'\widehat{\Sigma}_n(Y_n)W$. In other words, the estimate based on the rotated data equals the rotation of the estimate based on the original data. In the absence of any *a priori* knowledge about the structure of $\Sigma_n$, such as sparseness or a factor model, it is natural to consider only estimators of $\Sigma_n$ that are rotation-equivariant.

The class of rotation-equivariant estimators of the covariance that are a function of the sample covariance matrix is constituted of all the estimators that have the same eigenvectors as the sample covariance matrix; for example, see Perlman (2007, Section 5.4). Every such rotation-equivariant estimator is thus of the form

$$U_nD_nU_n' \quad \text{where} \quad D_n := \mathsf{Diag}(d_1, \ldots, d_p) \text{ is diagonal} , \tag{3.1}$$

and where $U_n$ is the matrix whose $i$th column is the sample eigenvector $u_i := u_{n,i}$. This is the class of rotation-equivariant estimators already studied by Stein (1975, 1986).

We can rewrite the expression for such a rotation-equivariant estimator as

$$U_nD_nU_n' = \sum_{i=1}^{p} d_i \cdot u_iu_i' . \tag{3.2}$$

This alternative expression shows that any such rotation-equivariant estimator is a linear combination of $p$ rank-1 matrices $u_iu_i'$ $(i = 1, \ldots, p)$. But since the $\{u_i\}$ form an orthonormal basis

in $\mathbb{R}^p$, the resulting estimator is still of full rank $p$, provided that all the weights $d_i$ $(i = 1, \ldots, p)$ are strictly positive.

**Remark 3.1** (Rotation-equivariant Estimators versus Structured Estimators)**.** By construction, the class (3.1) of rotation-invariant estimators have the same eigenvectors as the sample covariance matrix. In particular, consistent estimation of the covariance matrix is not possible under large-dimensional asymptotics.

Another approach would be to impose additional structure on the estimation problem, such as sparseness (Bickel and Levina, 2008), a graph model (Rajaratnam et al., 2008), or an (approximate) factor model (Fan et al., 2008).[1] The advantage of doing so is that, under suitable regularity conditions, consistent estimation of the covariance matrix is possible. The disadvantage is that if the assumed structure is misspecified, the estimator of the covariance matrix can be arbitrarily bad; and whether the structure is correctly specified can never be verified from the data alone.

Rotation-equivariant estimators are widely and successfully used in practice in situations where knowledge on additional structure is not available (or doubtful). This is evidenced by the many citations to Ledoit and Wolf (2004) who propose a linear shrinkage estimator that also belongs to the class (3.1); for example, see the beginning of Section 5.2. Therefore, developing a new, nonlinear shrinkage estimator that outperforms this previous proposal will be of substantial interest to applied researchers. ■

The first objective is to find the matrix in the class (3.1) of rotation-equivariant estimators that is closest to $\Sigma_n$. To measure distance, we choose the Frobenius norm defined as

$$||A||_F \coloneqq \sqrt{\mathsf{Tr}(AA')/r} \quad \text{for any matrix } A \text{ of dimension } r \times m \ . \tag{3.3}$$

(Dividing by the dimension of the square matrix $AA'$ inside the root is not standard, but we do this for asymptotic purposes so that the Frobenius norm remains constant equal to one for the identity matrix regardless of the dimension; see Ledoit and Wolf (2004).) As a result, we end up with the following minimization problem:

$$\min_{D_n} ||U_n D_n U_n' - \Sigma_n||_F \ .$$

Elementary matrix algebra shows that its solution is

$$D_n^* \coloneqq \mathsf{Diag}(d_1^*, \ldots, d_p^*) \quad \text{where} \quad d_i^* \coloneqq u_i' \Sigma_n u_i \ \text{ for } i = 1, \ldots, p \ . \tag{3.4}$$

Let $y \in \mathbb{R}^p$ be a random vector with covariance matrix $\Sigma_n$, drawn independently from the sample covariance matrix $S_n$. We can think of $y$ as an out-of-sample observation. Then $d_i^*$ is recognized as the variance of the linear combination $u_i'y$, conditional on $S_n$. In view of the expression (3.2), it makes intuitive sense that the matrices $u_i u_i'$ whose associated linear combination $u_i'y$ have higher out-of-sample variance should receive higher weight in computing the estimator of $\Sigma_n$.

[1] We only give one representative reference for each field here to save space.

The finite-sample optimal estimator is thus given by

$$S_n^* \coloneqq U_n D_n^* U_n' \quad \text{where} \quad D_n^* \text{ is defined as in (3.4)} . \tag{3.5}$$

(Clearly $S_n^*$ is not a *feasible* estimator, as it is based on the population covariance matrix $\Sigma_n$.)

By generalizing the Marčenko-Pastur equation (2.5), Ledoit and Péché (2011) show that $d_i^*$ can be approximated by the asymptotic quantities

$$d_i^{or} \coloneqq \begin{cases} \dfrac{1}{(c-1)\,\breve{m}_{\underline{F}}(0)} , & \text{if } \lambda_i = 0 \text{ and } c > 1 \\ \dfrac{\lambda_i}{\left|1 - c - c\,\lambda_i\,\breve{m}_F(\lambda_i)\right|^2} , & \text{otherwise} \end{cases} \qquad \text{for } i = 1, \ldots, p , \tag{3.6}$$

from which they deduce their oracle estimator

$$S_n^{or} \coloneqq U_n D_n^{or} U_n' \quad \text{where} \quad D_n^{or} \coloneqq \mathsf{Diag}(d_1^{or}, \ldots, d_p^{or}) . \tag{3.7}$$

The key difference between $D_n^*$ and $D_n^{or}$ is that the former depends on the unobservable population covariance matrix, whereas the latter depends on the limiting distribution of sample eigenvalues, $F$, which makes it amenable to consistent estimation. It turns out that this estimation problem is solved if a consistent estimator of the population eigenvalues $\boldsymbol{\tau}_n$ is available.

## 3.2 Nonlinear Shrinkage Estimator

### 3.2.1 The Case $p < n$

We start with the case $p < n$, which was already considered by Ledoit and Wolf (2012). Silverstein and Choi (1995) show how the support of $F$, denoted by $\mathsf{Supp}(F)$, is determined; also see Section 2.3 of Ledoit and Wolf (2012). $\mathsf{Supp}(F)$ is seen to be the union of a finite number of disjoint compact intervals, bounded away from zero. To simplify the discussion, we will assume from here on that $\mathsf{Supp}(F)$ is a single compact interval, bounded away from zero, with $F' > 0$ in the interior of this interval. But if $\mathsf{Supp}(F)$ is the union of a finite number of such intervals, the arguments presented in this section as well as in the remainder of the paper apply separately to each interval. In particular, our consistency results presented below can be easily extended to this more general case.

Recall that, for any $\mathbf{t} \coloneqq (t_1, \ldots, t_p)' \in [0, +\infty)^p$, equations (2.13)–(2.14) define $m_{n,p}^{\mathbf{t}}$ as the Stieltjes transform of $F_{n,p}^{\mathbf{t}}$, the limiting distribution of sample eigenvalues corresponding to the population spectral distribution $p^{-1} \sum_{i=1}^p \mathbb{1}_{[t_i, +\infty)}$. Its domain is the strict upper half of the complex plane, but it can be extended to the real line since Silverstein and Choi (1995) prove that $\forall \lambda \in \mathbb{R} - \{0\} \quad \lim_{z \in \mathbb{C}^+ \to \lambda} m_{n,p}^{\mathbf{t}}(z) =: \breve{m}_{n,p}^{\mathbf{t}}(\lambda)$ exists.

Ledoit and Wolf (2012) show how a consistent estimator of $\breve{m}_F$ can be derived from a consistent estimator of $\boldsymbol{\tau}_n$, such as $\widehat{\boldsymbol{\tau}}_n$ defined in Theorem 2.2. Their Proposition 4.3 establishes that $\breve{m}_{n,p}^{\widehat{\boldsymbol{\tau}}_n}(\lambda) \to \breve{m}_F(\lambda)$ uniformly in $\lambda \in \mathsf{Supp}(F)$, except for two arbitrarily small regions at the

lower and upper end of $\mathsf{Supp}(F)$. Replacing $\breve{m}_F$ with $\breve{m}_{n,p}^{\widehat{\boldsymbol{\tau}}_n}$ and $c$ with $p/n$ in Ledoit and Péché (2011)'s oracle quantities $d_i^{or}$ of (3.6) yields

$$\widehat{d}_i \coloneqq \frac{\lambda_i}{\left|1 - \frac{p}{n} - \frac{p}{n}\,\lambda_i \cdot \breve{m}_{n,p}^{\widehat{\boldsymbol{\tau}}_n}(\lambda_i)\right|^2} \quad \text{for } i = 1, \ldots, p\ . \tag{3.8}$$

(Note here that in the case $p < n$, all sample eigenvalues $\lambda_i$ are positive almost surely, for $n$ large enough, by the results of Bai and Silverstein (1998).) In turn, the *bona fide* nonlinear shrinkage estimator of $\Sigma_n$ is obtained as:

$$\widehat{S}_n \coloneqq U_n \widehat{D}_n U_n' \quad \text{where} \quad \widehat{D}_n \coloneqq \mathsf{Diag}(\widehat{d}_1, \ldots, \widehat{d}_p)\ . \tag{3.9}$$

### 3.2.2 The Case $p > n$

We move on to the case $p > n$, which was not considered by Ledoit and Wolf (2012). In this case, $F$ is a mixture distribution with a discrete part and a continuous part. The discrete part is a point mass at zero with mass $(c-1)/c$. The continuous part has total mass $1/c$ and its support is the union of a finite number of disjoint intervals, bounded away from zero; again, see Silverstein and Choi (1995).

It can be seen from equations (2.6)–(2.9) that $\underline{F}$ corresponds to the continuous part of $F$, scaled to be a proper distribution (function): $\lim_{t\to\infty} \underline{F}(t) = 1$. Consequently, $\mathsf{Supp}(F) = \{0\} \cup \mathsf{Supp}(\underline{F})$. To simplify the discussion, we will assume from here on that $\mathsf{Supp}(\underline{F})$ is a single compact interval, bounded away from zero, with $\underline{F}' > 0$ in the interior of this interval. But if $\mathsf{Supp}(\underline{F})$ is the union of a finite number of such intervals, the arguments presented in this section as well as in the remainder of the paper apply separately to each interval. In particular, our consistency results presented below can be easily extended to this more general case.

The oracle quantities $d_i^{or}$ of (3.6) involve $\breve{m}_{\underline{F}}(0)$ and $\breve{m}_F(\lambda_i)$ for various $\lambda_i > 0$; recall that $\breve{m}_{\underline{F}}(0)$ exists in the case $c > 1$.

Using the original Marčenko-Pastur equation (2.10), a strongly consistent estimator of the quantity $\breve{m}_{\underline{F}}(0)$ is the unique solution $m \coloneqq \widehat{\breve{m}_{\underline{F}}(0)}$ in $(0, \infty)$ to the equation

$$m = \left[\frac{1}{n}\sum_{i=1}^{p} \frac{\widehat{\tau}_i}{1 + \widehat{\tau}_i\, m}\right]^{-1}, \tag{3.10}$$

where $\widehat{\boldsymbol{\tau}}_n \coloneqq (\widehat{\tau}_1, \ldots, \widehat{\tau}_p)'$ is defined as in Theorem 2.2.

Again, since $\widehat{\boldsymbol{\tau}}_n$ is consistent for $\boldsymbol{\tau}_n$, Proposition 4.3 of Ledoit and Wolf (2012) implies that $\breve{m}_{n,p}^{\widehat{\boldsymbol{\tau}}_n}(\lambda) \to \breve{m}_F(\lambda)$ uniformly in $\lambda \in \mathsf{Supp}(\underline{F})$, except for two arbitrarily small regions at the lower and upper end of $\mathsf{Supp}(\underline{F})$.

Finally, the *bona fide* nonlinear shrinkage estimator of $\Sigma_n$ is obtained as (3.9) but now with

$$\widehat{d}_i \coloneqq \begin{cases} \dfrac{\lambda_i}{\left|1 - \frac{p}{n} - \frac{p}{n}\,\lambda_i \cdot \breve{m}_{n,p}^{\widehat{\boldsymbol{\tau}}_n}(\lambda_i)\right|^2}\ , & \text{if } \lambda_i > 0 \\[2ex] \dfrac{1}{\left(\frac{p}{n} - 1\right)\,\widehat{\breve{m}_{\underline{F}}(0)}}\ , & \text{if } \lambda_i = 0 \end{cases} \qquad \text{for } i = 1, \ldots, p\ , \tag{3.11}$$

### 3.3 Strong Consistency

The following theorem establishes that our nonlinear shrinkage estimator, based on the estimator $\widehat{\boldsymbol{\tau}}_n$ of Theorem 2.2, is strongly consistent for the oracle estimator across both cases $p < n$ and $p > n$.

**Theorem 3.1.** *Let $\widehat{\boldsymbol{\tau}}_n$ be an estimator of the eigenvalues of the population covariance matrix satisfying $p^{-1} \sum_{i=1}^{p} [\widehat{\tau}_{n,i} - \tau_{n,i}]^2 \xrightarrow{\text{a.s.}} 0$. Define the nonlinear shrinkage estimator $\widehat{S}_n$ as in (3.9), where the $\widehat{d}_i$ are as in (3.8) in the case $p < n$ and as in (3.11) in the case $p > n$.*

*Then $||\widehat{S}_n - S_n^{or}||_F \xrightarrow{\text{a.s.}} 0$.*

**Remark 3.2.** We have to rule out the case $c = 1$ (or $p = n$) for mathematical reasons.

First, we need $\mathsf{Supp}(\underline{F})$ to be bounded away from zero to establish various consistency results. But when $c = 1$, then $\mathsf{Supp}(\underline{F})$ can start at zero, that is, there exists $u > 0$ such that $F'(\lambda) > 0$ for all $\lambda \in (0, u)$. This was already established by Marčenko and Pastur (1967) for the special case when $H$ is a point mass at one. In particular, the resulting (standard) Marčenko-Pastur distribution $F$ has density function

$$F'(\lambda) = \begin{cases} \frac{1}{2\pi\lambda c}\sqrt{(b-\lambda)(\lambda-a)}\ , & \text{if } a \leq \lambda \leq b\ , \\ 0\ , & \text{otherwise}\ , \end{cases}$$

and has point mass $(c-1)/c$ at the origin if $c > 1$, where $a := (1-\sqrt{c})^2$ and $b := (1+\sqrt{y})^2$; for example, see Bai and Silverstein (2010, Section 3.3.1).

Second, we also need the assumption $c \neq 1$ 'directly' in the proof of Theorem 3.1 to demonstrate that the summand $D_1$ in (A.17) converges to zero.

Although the case $c = 1$ is not covered by the mathematical treatment, we can still address it in Monte Carlo simulations; see Section 5.2. ■

## 4 Principal Component Analysis

Principal component analysis (PCA) is one of the oldest and best-known techniques of multivariate analysis, dating back to Pearson (1901) and Hotelling (1933); for a comprehensive treatment, see Jolliffe (2002).

### 4.1 The Central Idea and the Common Practice

The central idea of PCA is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming the original variables to a new set of *uncorrelated* variables, the principal components, which are ordered so that the 'largest' *few* retain most of the variation present in *all* of the original variables.

More specifically, let $y \in \mathbb{R}^p$ be a random vector with covariance matrix $\Sigma$; in this section, it will be convenient to drop the subscript $n$ from the covariance matrix and related quantities. Let $((\tau_1, \ldots, \tau_p); (v_1, \ldots, v_p))$ denote a system of eigenvalues and eigenvectors of $\Sigma$. To be

consistent with our former notation, we assume that the eigenvalues $\tau_i$ are sorted in *increasing* order. Then the principal components of $y$ are given by $v_1'y, \ldots, v_p'y$. Since the eigenvalues $\tau_i$ are sorted in increasing order, the principal component with the largest variance is $v_p'y$ and the principal component with the smallest variance is $v_1'y$. The eigenvector $v_i$ is called the *vector of coefficients* or *loadings* for the $i$th principal component ($i = 1, \ldots, p$).

Two brief remarks are in order. First, some authors use the term *principal components* for the eigenvectors $v_i$; but we agree with Jolliffe (2002, Section 1.1) that this usage is confusing and that it is preferable to reserve the term for the derived variables $v_i'y$. Second, in the PCA literature, in contrast to the bulk of the multivariate statistics literature, the eigenvalues $\tau_i$ are generally sorted in decreasing order so that $v_1'y$ is the 'largest' principal component (that is, the principal component with the largest variance). This is understandable when the goal is expressed as capturing most of the total variation in the *first* few principal components. But to avoid confusion with other sections of this paper, we keep the convention of eigenvalues being sorted in increasing order, and then express the goal as capturing most of the total variation in the *largest* few principal components.

The $k$ largest principal components in our notation are thus given by $v_p'y, \ldots, v_{p-k+1}'y$ ($k = 1, \ldots, p$). Their (cumulative) fraction captured of the total variation contained in $y$, denoted by $f_k(\Sigma)$, is given by

$$f_k(\Sigma) = \frac{\sum_{j=1}^{k} \tau_{p-j+1}}{\sum_{m=1}^{p} \tau_m} \ , \quad k = 1, \ldots, p \ . \tag{4.1}$$

The most common rule in deciding how many principal components to retain is to decide on a given fraction of the total variation that one wants to capture, denoted by $f_{target}$, and to then retain the largest $k$ principal components, where $k$ is the smallest integer satisfying $f_k(\Sigma) \geq f_{target}$. Commonly chosen values of $f_{target}$ are $70\%, 80\%, 90\%$, depending on the context. For obvious reasons, this rule is known as the *cumulative-percentage-of-total-variation* rule.

There exist a host of other rules, either analytical or graphical, such as Kaiser's rule or the scree plot; see Jolliffe (2002, Section 6.1). The vast majority of these rules are also solely based on the eigenvalues $(\tau_1, \ldots, \tau_p)$.

The problem is that generally the covariance matrix $\Sigma$ is unknown. Thus, neither the (population) principal components $v_i'y$ nor their cumulative percentages of total variation $f_k(\Sigma)$ can be used in practice.

The common solution is to replace $\Sigma$ with the sample covariance matrix $S$, computed from a random sample $y_1, \ldots, y_n$, independent of $y$. Let $((\lambda_1, \ldots, \lambda_p); (u_1, \ldots, u_p))$ denote a system of eigenvalues and eigenvectors of $S$; it is assumed again that the eigenvalues $\lambda_i$ are sorted in increasing order. Then the (sample) principal components of $y$ are given by $u_1'y, \ldots, u_p'y$.

The various rules in deciding how many (sample) principal components to retain are now based on the sample eigenvalues $\lambda_i$. For example, the cumulative-percentage-of-total-variation rule retains the largest $k$ principal components, where $k$ is the smallest integer satisfying

$f_k(S) \geq f_{target}$, with

$$f_k(S) = \frac{\sum_{j=1}^{k} \lambda_{p-j+1}}{\sum_{m=1}^{p} \lambda_m} \ , \quad k = 1, \ldots, p \ . \tag{4.2}$$

The pitfall in doing so, unless $p \ll n$, is that $\lambda_i$ is not a good estimator of the variance of the $i$th principal component. Indeed, the variance of the $i$th principal component, $u_i'y$, is given by $u_i'\Sigma u_i$ rather than by $\lambda_i = u_i'Su_i$. By design, for large values of $i$, the estimator $\lambda_i$ is upward biased for the true variance $u_i'\Sigma u_i$, whereas for small values of $i$, it is downward biased. In other words, the variances of the large principal components are overestimated whereas the variances of the small principal components are underestimated. The unfortunate consequence is that most rules in deciding how many principal components to retain, such as the cumulative-percentage-of-total-variation rule, generally retain fewer principal components than really needed.

## 4.2 Previous Approaches under Large-Dimensional Asymptotics

All the previous approaches under large-dimensional asymptotics that we are aware of impose some additional structure on the estimation problem.

Most works assume a sparseness conditions on the eigenvectors $v_i$ or on the covariance matrix $\Sigma$; see Amini (2011) for a comprehensive review.

Mestre (2008), on the other hand and as discussed before, assumes that $\Sigma$ has only $\bar{M} \ll p$ distinct eigenvalues and further that the multiplicity of each of the $\bar{M}$ distinct eigenvalues is known (which implies that the number $\bar{M}$ is known as well). Furthermore, he needs spectral separation. In this restrictive setting, he is able to construct a consistent estimator of every distinct eigenvalue and its associated eigenspace (that is, the space spanned by all eigenvectors corresponding to a specific distinct eigenvalue).

## 4.3 Alternative Approach Based on Nonlinear Shrinkage

Unlike previous approaches under large-dimensional asymptotics, we do not wish to impose additional structure on the estimation problem. As mentioned before, in such a rotation-equivariant setting, consistent (or even improved) estimators of the eigenvectors $v_i$ are not available and one must indeed use the sample eigenvectors $u_i$ as the loadings. Therefore, as is common practice, the principal components used are the $u_i'y$.

Ideally, the rules in deciding how many principal components to retain should be based on the variances of the principal components given by $u_i'\Sigma u_i$. It is important to note that even if the population eigenvalues $\tau_i$ were known, the rules should not be based on them. This is because the population eigenvalues $\tau_i = v_i'\Sigma v_i$ describe the variances of the $v_i'y$, which are not available and thus not used. It seems that this important point has not been realized so far. Indeed, various authors have used PCA as a motivational example in the estimation of the population eigenvalues $\tau_i$; for example, see El Karoui (2008), Mestre (2008), and Yao et al. (2012). But unless the population eigenvectors $v_i$ are available as well, using the $\tau_i$ is misleading.

Although, in the absence of additional structure, it is not possible to construct improved principal components, it is possible to accurately estimate the variances of the commonly-used principal components. This is because the variance of the $i$th principal component is nothing else than the finite-sample-optimal nonlinear shrinkage constant $d_i^*$; see equation (3.4). Its oracle counterpart $d_i^{or}$ is given in equation (3.6) and the *bona fide* counterpart $\widehat{d}_i$ is given in equation (3.8) in the case $p < n$ and in equation (3.11) in the case $p > n$.

Our solution then is to base the various rules in deciding how many principal components to retain on the $\widehat{d}_i$ in place of the unavailable $d_i^* = u_i' \Sigma u_i$. For example, the cumulative-percentage-of-total-variation rule retains the $k$ largest principal components, where $k$ is the smallest integer satisfying $f_k(\widehat{S}) \geq f_{target}$, with

$$f_k(\widehat{S}) = \frac{\sum_{j=1}^{k} \widehat{d}_{p-j+1}}{\sum_{m=1}^{p} \widehat{d}_m} \ , \quad k = 1, \ldots, p \ . \tag{4.3}$$

**Remark 4.1.** We have taken the total variation to be $\sum_{m=1}^{p} d_m^*$, and the variation attributable to the $k$ largest principal components to be $\sum_{j=1}^{k} d_{p-j+1}^*$. In general, the sample principal components $u_i' y$ are not uncorrelated (unlike the population principal components $v_i' y$). This means that $u_i' \Sigma u_j$ can be non-zero for $i \neq j$. Nonetheless, even in this case, the variation attributable to the $k$ largest principal components is still equal to $\sum_{j=1}^{k} d_{p-j+1}^*$, as explained in Appendix B. ■

While most applications of PCA seek the principal components with the largest variances, there are also some applications of PCA that seek the principal components with the *smallest* variances; see Jolliffe (2002, Section 3.4). In the case $p > n$, a certain number of the $\lambda_i$ will be equal to zero, falsely giving the impression that a certain number of the smallest principal components have variance zero. Such applications also highlight the use of replacing the $\lambda_i$ with our nonlinear shrinkage constants $\widehat{d}_i$, which are always greater than zero.

# 5 Monte Carlo Simulations

In this section, we study the finite-sample performance of various estimators in different settings.

## 5.1 Estimation of Population Eigenvalues

We first focus on estimating the eigenvalues of the population covariance matrix, $\boldsymbol{\tau}_n$. Of major interest to us is the case where all the eigenvalues are or can be distinct; but we also consider the case where they are known or assumed to be grouped into a small number of high-multiplicity clusters.

### 5.1.1 All Distinct Eigenvalues

We consider the following estimators of $\boldsymbol{\tau}_n$.

- **Sample:** The sample eigenvalues $\lambda_{n,i}$.
- **Lawley:** The bias-corrected sample eigenvalues using the formula of Lawley (1956, Section 4). This transformation may not be monotonic in finite samples. Therefore, we post-process it with an isotonic regression.
- **El Karoui:** The estimator of El Karoui (2008). It provides an estimator of $H_n$, not $\boldsymbol{\tau}_n$, so we derive estimates of the population eigenvalues using 'smoothed' quantiles in the spirit of equations (2.17)–(2.16).[2]
- **LW:** Our estimator $\widehat{\boldsymbol{\tau}}_n$ of Theorem 2.2.

It should be pointed out that the estimator of Lawley (1956) is designed to reduce the finite-sample bias of the sample eigenvalues $\lambda_{n,i}$; it is not necessarily designed for consistent estimation of $\boldsymbol{\tau}_n$ under large-dimensional asymptotics.

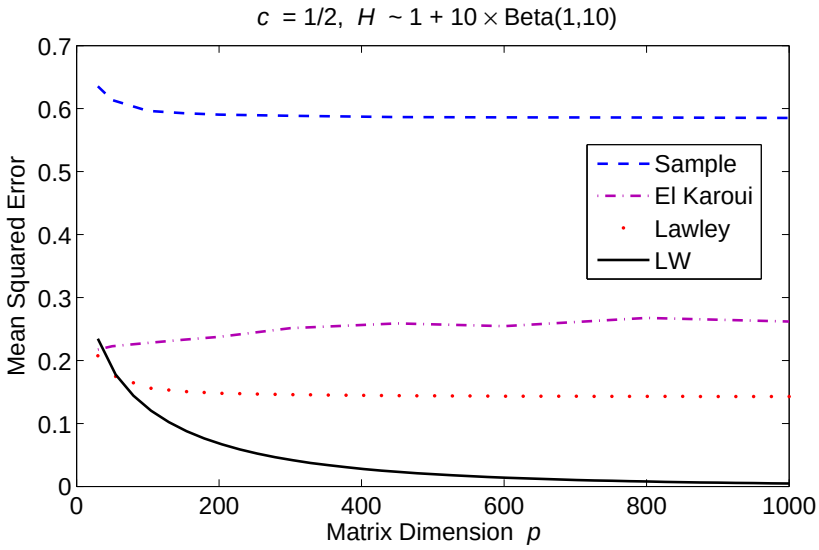
Let $\widetilde{\tau}_{n,i}$ denote a generic estimator of $\tau_{n,i}$. The evaluation criterion is the dimension-normalized Euclidian distance between estimated eigenvalues $\widetilde{\boldsymbol{\tau}}_n$ and population eigenvalues $\boldsymbol{\tau}_n$:

$$\frac{1}{p}\sum_{i=1}^{p}\left[\widetilde{\tau}_{n,i}-\tau_{n,i}\right]^2 \quad , \tag{5.1}$$

averaged over 1,000 Monte Carlo simulations in each scenario.

Convergence

In the first design, the $i$th population eigenvalue is equal to $\tau_{n,i} := H^{-1}((i-0.5)/p)$ $(i=1,\ldots,p)$, where $H$ is given by the distribution of $1+10W$, and $W \sim \text{Beta}(1,10)$; this distribution is right-skewed and resembles in shape an exponential distribution. The distribution of the random variates comprising the $n \times p$ data matrix $X_n$ is real Gaussian. We fix the concentration at $p/n = 0.5$ and vary the dimension from $p = 30$ to $p = 1,000$. The results are displayed in Figure 2.




[2]We implemented this estimator to the best of our abilities, following the description in El Karoui (2008). Despite several attempts, we were not able to obtain the original code.

Figure 2: Convergence of estimated eigenvalues to population eigenvalues in the case where the sample covariance matrix is nonsingular.

It can be seen that the empirical mean squared error for LW converges to zero, which is in agreement with the proven consistency of Theorem 2.2. For all the other estimators, the average distance from $\boldsymbol{\tau}_n$ appears bounded away from zero. This simulation also shows that dividing by $p$ is indeed the appropriate normalization for the Euclidian norm in equation (2.18), as it drives a wedge between estimators such as the sample eigenvalues that are not consistent and $\widehat{\boldsymbol{\tau}}_n$, which is consistent.

The second design is similar to the first design, except that we fix the concentration at $p/n = 2$ and now vary the sample size from $n = 30$ to $n = 1,000$ instead of the dimension. In this design, the sample covariance matrix is always singular. The results are displayed in Figure 3 and are qualitatively similar. Again, LW is the only estimator that appears to be consistent. Notice the vertical scale: the difference between El Karoui and LW is of the same order of magnitude as in Figure 2, but Sample and Lawley are vastly more erroneous now.

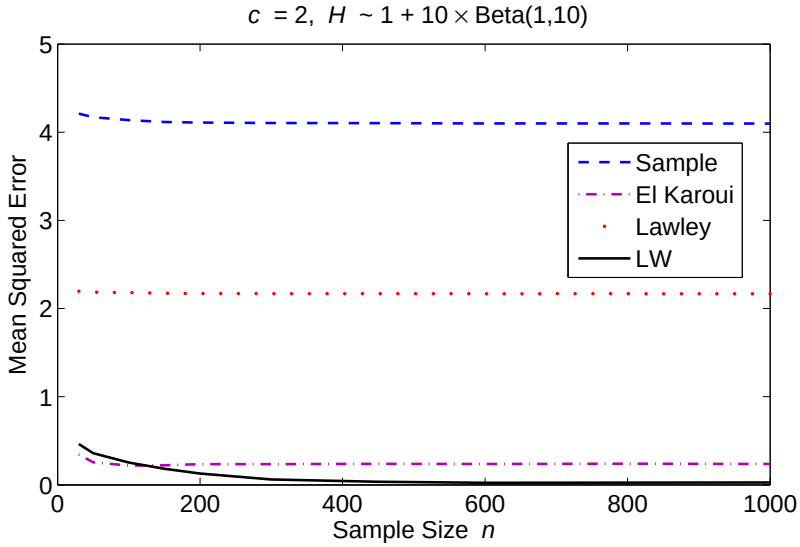



Figure 3: Convergence of estimated eigenvalues to population eigenvalues in the case where the sample covariance matrix is singular.

### Condition Number

In the third design, the focus is on the condition number. The $i$th population eigenvalue is still $\tau_{n,i} := H^{-1}((i-0.5)/p)$ $(i = 1, \ldots, p)$, but $H$ is now given by the distribution of $a + 10W$, where $W \sim \text{Beta}(1, 10)$, and $a \in [0, 7]$. As a result, the smallest eigenvalue approaches $a$, and the previously-used distribution for $H$ is included as a special case when $a = 1$. The condition number decreases in $a$ from approximately $10,000$ to 2.4.

We use $n = 1,600$ and $p = 800$, so that $p/n = 0.5$. The random variates are still real Gaussian. The results are displayed in Figure 4. It can be seen that Sample and Lawley perform quite well for values of $a$ near zero (that is, for very large condition numbers) but their performance gets worse as $a$ increases (that is, as the condition number decreases). On

the other hand, the performance of El Karoui is more stable across all values of $a$, though relatively bad. The performance of LW is uniformly the best and also stable across $a$.

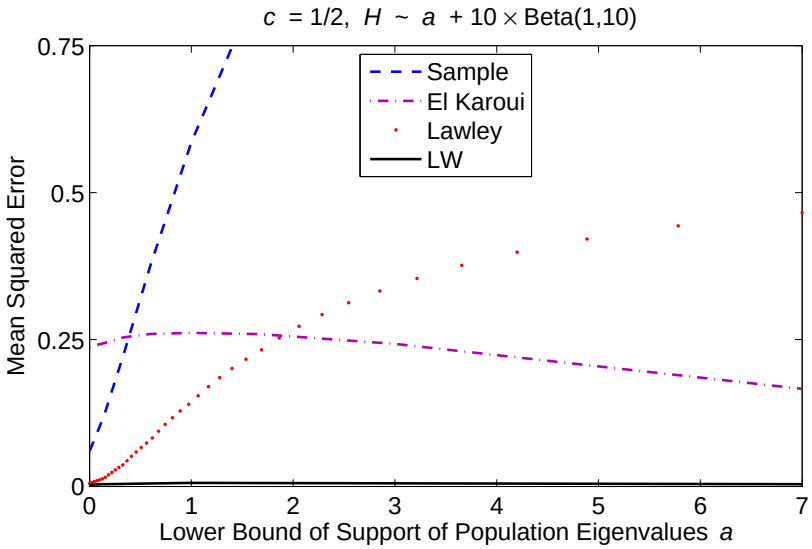



Figure 4: Effect of the condition number on the mean squared error between estimated and population eigenvalues.

Shape of the Distribution

In the fourth design, we consider a wide variety of shapes of $H$, which is now given by the distribution of $1+10W$, where $W$ follows a Beta distribution with parameters $\{(1,1),(1,2),(2,1),(1.5,1.5),(0.5,0.5),(5,5),(5,2),(2,5)\}$; see Figure 7 of Ledoit and Wolf (2012) for a graphical representation of the corresponding densities. Always again, the $i$th population eigenvalue is $\tau_{n,i} := H^{-1}((i-0.5)/p)$ $(i=1,\ldots,p)$.

We use $n=1,600$ and $p=800$, so that $p/n=0.5$. The random variates are real Gaussian. The results are presented in Table 1. It can be seen that LW is uniformly best and Sample is uniformly worst. There is no clear-cut ranking for the remaining two estimators. On average, Lawley is second best, followed by El Karoui.

| Parameters | LW | Sample | El Karoui | Lawley |
|---|---|---|---|---|
| $(1,1)$ | 0.15 | 6.70 | 2.65 | 0.66 |
| $(1,2)$ | 0.06 | 2.58 | 1.65 | 0.27 |
| $(2,1)$ | 0.16 | 15.59 | 2.23 | 2.61 |
| $(1.5,1.5)$ | 0.09 | 7.07 | 2.03 | 0.93 |
| $(0.5,0.5)$ | 0.08 | 7.04 | 2.87 | 0.53 |
| $(5,5)$ | 0.08 | 9.52 | 1.02 | 2.13 |
| $(5,2)$ | 0.12 | 20.93 | 1.39 | 4.90 |
| $(2,5)$ | 0.08 | 2.59 | 0.87 | 0.46 |
| Average | 0.10 | 9.00 | 1.84 | 1.56 |

Table 1: Mean squared error between estimated and population eigenvalues.

Heavy Tails

So far, the variates making up the data matrix $X_n$ always had a Gaussian distribution. It is also of interest to consider a heavy-tailed distribution instead. We return to the first design with $n = 1600$ and $p = 800$, so that $p/n = 0.5$. In addition to the Gaussian distribution, which can be viewed as a $t$-distribution with infinite degrees of freedom, we also consider a the $t$-distribution with three degrees of freedom (scaled to have unit variance). The results are presented in Table 2. It can be seen that all estimators perform worse when the degrees of freedom are changed from infinity to three, but LW is still by far the best.

| Degrees of Freedom | LW | Sample | El Karoui | Lawley |
|---|---|---|---|---|
| 3 | 0.21 | 4.97 | 4.02 | 4.41 |
| $\infty$ | 0.01 | 0.59 | 0.27 | 0.14 |

Table 2: Mean squared error between estimated and population eigenvalues.

### 5.1.2 Clustered Eigenvalues

We are mainly interested in the case where the population eigenvalues are or can be distinct, but it is also worthwhile seeing how (an adapted version of) our estimator of $\boldsymbol{\tau}_n$ compares to the one of Mestre (2008) in the setting where the population eigenvalues are known or assumed to be grouped into a small number of high-multiplicity clusters.

Let $\gamma_1 < \gamma_2 < \cdots < \gamma_{\bar{M}}$ denote the set of pairwise different eigenvalues of the population covariance matrix $\Sigma$, where $\bar{M}$ is the number of distinct population eigenvalues ($1 \leq \bar{M} < p$). Each of the eigenvalues $\gamma_j$ has known multiplicity $K_j$ ($j = 1, \ldots, \bar{M}$), so that $p = \sum_{j=1}^{\bar{M}} K_j$. (Knowing the multiplicities of the eigenvalues $\gamma_j$ comes from knowing their masses $m_j$ in the limiting spectral distribution $H$, as assumed in Mestre (2008): $K_j/p = m_j$.)

Then the optimization problem in Theorem 2.1 becomes:

$$\widehat{\boldsymbol{\gamma}}_n \coloneqq \underset{(\gamma_1,\gamma_2,\ldots,\gamma_{\bar{M}})\in[0,\infty)^{\bar{M}}}{\operatorname{argmin}} \frac{1}{p}\sum_{i=1}^{p}\left[\lambda_{n,i} - q_{n,p}^{i}(\mathbf{t})\right]^2 \tag{5.2}$$

$$\text{subject to:} \quad \mathbf{t} = (\underbrace{\gamma_1,\ldots,\gamma_1}_{K_1 \text{ times}}, \underbrace{\gamma_2,\ldots,\gamma_2}_{K_2 \text{ times}}, \ldots, \underbrace{\gamma_{\bar{M}},\ldots,\gamma_{\bar{M}}}_{K_{\bar{M}} \text{ times}})' \tag{5.3}$$

$$\gamma_1 < \gamma_2 < \ldots < \gamma_{\bar{M}} \tag{5.4}$$

We consider the following estimators of $\boldsymbol{\tau}_n$.

- **Traditional:** $\gamma_j$ is estimated by the average of all corresponding sample eigenvalues $\lambda_{n,i}$; under the condition of spectral separation assumed in Mestre (2008), it is known which $\gamma_j$ corresponds to which $\lambda_{n,i}$.

- **Mestre:** The estimator defined in Mestre (2008, Theorem 3).
- **LW:** Our modified estimator as defined in (5.2)–(5.4).

The mean squared error criterion (5.1) specializes in this setting to

$$\sum_{j=1}^{\bar{M}} m_j \left(\widehat{\gamma}_j - \gamma_j\right)^2 .$$

We report the average MSE over 1,000 Monte Carlo simulations in each scenario.

Convergence

The first design is based on Tables I and II of Mestre (2008). The distinct population eigenvalues are $(\gamma_1, \gamma_2, \gamma_3, \gamma_4) = (1, 7, 15, 25)$ with respective multiplicities $(K_1, K_2, K_3, K_4) = (p/2, p/4, p/8, p/8)$. The distribution of the random variates comprising the $n \times p$ data matrix $X_n$ is circular symmetric complex Gaussian, as in Mestre (2008). We fix the concentration at $p/n = 0.32$ and vary the dimension from $p = 8$ to $p = 1,000$; the lower end $p = 8$ corresponds to Table I in Mestre (2008), while the upper end $p = 1,000$ corresponds to Table II in Mestre (2008). The results are displayed in Figure 5. It can be seen that the average MSE of both Mestre and LW converges to zero, and that the performance of the two estimators is nearly indistinguishable. On the other hand, Traditional is seen to be inconsistent, as its MSE remains bounded away from zero.

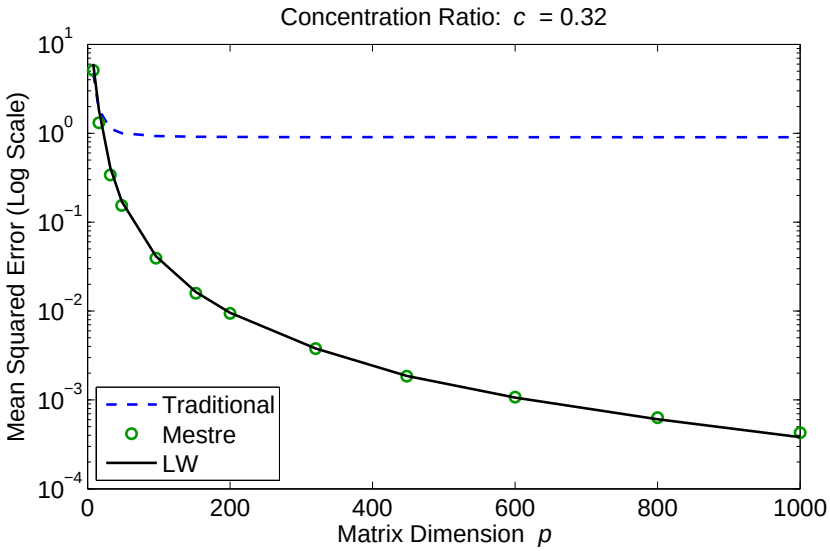



Figure 5: Convergence of estimated eigenvalues to population eigenvalues when eigenvalues are grouped into a small number of high-multiplicity clusters.

Performance When One Eigenvalue Is Isolated

The second design is based on Table III of Mestre (2008). The distinct population eigenvalues are $(\gamma_1, \gamma_2, \gamma_3, \gamma_4) = (1, 7, 15, 25)$ with multiplicities $(K_1, K_2, K_3, K_4) = (160, 80, 79, 1)$. There is a single 'isolated' large eigenvalue. The distribution of the random variates comprising the $n \times p$ data matrix $X_n$ is circular symmetric complex Gaussian, as in Mestre (2008). We use

$n = 1,000$ and $p = 320$, so that $p/n = 0.32$. The averages and the standard deviations of the estimates $\widehat{\gamma}_j$ over 10,000 Monte Carlos simulations are presented in Table 3; note here that the numbers for Traditional and Mestre have been directly copied from Table III of Mestre (2008). The inconsistency of Traditional is again apparent. In terms of estimating $(\gamma_1, \gamma_2, \gamma_3)$, the performance of Mestre and LW is nearly indistinguishable. In terms of estimating $\gamma_4$, Mestre has a smaller bias (in absolute value) while LW has a smaller standard deviation; combining the two criteria yields a mean squared error of $(25 - 24.9892)^2 + 1.0713^2 = 1.1478$ for Mestre and a mean squared error of $(25 - 24.9238)^2 + 0.8898^2 = 0.7976$ for LW.

| | | Traditional | | Mestre | | LW | |
|---|---|---|---|---|---|---|---|
| Eigenvalue | Multiplicity | Mean | Std. Dev. | Mean | Std. Dev. | Mean | Std. Dev. |
| $\gamma_1 = 1$ | 160 | 0.8210 | 0.0023 | 0.9997 | 0.0032 | 1.0006 | 0.0034 |
| $\gamma_2 = 7$ | 80 | 6.1400 | 0.0208 | 6.9942 | 0.0343 | 7.0003 | 0.0319 |
| $\gamma_3 = 15$ | 79 | 16.1835 | 0.0514 | 14.9956 | 0.0681 | 14.9995 | 0.0580 |
| $\gamma_4 = 25$ | 1 | 28.9104 | 0.7110 | 24.9892 | 1.0713 | 24.9238 | 0.8898 |

Table 3: Empirical mean and standard deviation of the eigenvalue estimator of Mestre (2008), sample eigenvalues, and the proposed estimator. The first six columns are copied from Table III of Mestre (2008). Results are based on $10,000$ Monte Carlo simulations with circularly symmetric complex Gaussian random variates.

### 5.2 Covariance Matrix Estimation

As detailed in Section 3.1, the finite-sample optimal estimator in the class of rotation-equivariant estimators is given by $S_n^*$ as defined in (3.5). As the benchmark, we use the linear shrinkage estimator of Ledoit and Wolf (2004) instead of the sample covariance matrix. We do this because the linear shrinkage estimator has become the *de facto* standard among leading researchers because of its simplicity, accuracy, and good conditioning properties. It has been used in several fields of statistics, such as linear regression with a large number of regressors (Anatolyev, 2012), linear discriminant analysis (Pedro Duarte Silva, 2011), factor analysis (Lin and Bentler, 2012), unit root tests (Demetrescu and Hanck, 2012), and vector autoregressive models (Huang and Schneider, 2011), among others. Beyond pure statistics, the linear shrinkage estimator has been applied in finance for portfolio selection (Tsagaris et al., 2012) and tests of asset pricing models (Khan, 2008); in signal processing for cellular phone transmission (Nguyen et al., 2011) and radar detection (Wei et al., 2011); and in biology for neuroimaging (Varoquaux et al., 2010), genetics (Lin et al., 2012), cancer research (Pyeon et al., 2007), and psychiatry (Markon, 2010). It has also been used in such varied applications as physics (Pirkl et al., 2012), chemistry (Guo et al., 2012), climatology (Ribes et al., 2009), oil exploration (Sætrom et al., 2012), road safety research (Haufe et al., 2011), etc. In summary,

the comparatively poor performance of the sample covariance matrix and the popularity of the linear shrinkage estimator justify taking the latter as the benchmark.

The improvement of the nonlinear shrinkage estimator $\widehat{S}_n$ over the linear shrinkage estimator of Ledoit and Wolf (2004), denoted by $\overline{S}_n$, will be measured by how closely this estimator approximates the finite-sample optimal estimator $S_n^*$ relative to $\overline{S}_n$. More specifically, we report the Percentage Relative Improvement in Average Loss (PRIAL), which is defined as

$$\text{PRIAL} \coloneqq \text{PRIAL}(\widehat{\Sigma}_n) \coloneqq 100 \times \left\{ 1 - \frac{\mathbb{E}\left[\left\|\widehat{\Sigma}_n - S_n^*\right\|_F^2\right]}{\mathbb{E}\left[\left\|\overline{S}_n - S_n^*\right\|_F^2\right]} \right\} \% \ , \tag{5.5}$$

where $\widehat{\Sigma}_n$ is an arbitrary estimator of $\Sigma_n$. By definition, the PRIAL of $\overline{S}_n$ is 0% while the PRIAL of $S_n^*$ is 100%.

We consider the following estimators of $\Sigma_n$.

- **LW (2012) Estimator:** The nonlinear shrinkage estimator of Ledoit and Wolf (2012); this version only works for the case $p < n$.
- **New Nonlinear Shrinkage Estimator:** The new nonlinear shrinkage estimator of Section 3.2; this version works across both cases $p < n$ and $p > n$.
- **Oracle:** The (infeasible) oracle estimator of Section 3.1.

Convergence

In our design, 20% of the population eigenvalues are equal to 1, 40% are equal to 3, and 40% are equal to 10. This is a particularly interesting and difficult example introduced and analyzed in detail by Bai and Silverstein (1998); it has also been used in previous Monte Carlo simulations by Ledoit and Wolf (2012). The distribution of the random variates comprising the $n \times p$ data matrix $X_n$ is real Gaussian. We study convergence of the various estimators by keeping the concentration $p/n$ fixed while increasing the sample size $n$. We consider the three cases $p/n = 0.5, 1, 2$; as discussed in Remark 3.2, the case $p/n = 1$ is not covered by the mathematical treatment. The results are displayed in Figure 6, which shows empirical PRIAL's across 1,000 Monte Carlo simulations (one panel for each case $p/n = 0.5, 1, 2$). It can be seen that the new nonlinear shrinkage estimator always outperforms linear shrinkage with its PRIAL converging to 100%, though slower than the oracle estimator. As expected, the relative improvement over the linear shrinkage estimator is inversely related to the concentration ratio; also see Figure 4 of Ledoit and Wolf (2012). In the case $p < n$, it can also be seen that the new nonlinear shrinkage estimator slightly outperforms the earlier nonlinear shrinkage estimator of Ledoit and Wolf (2012). Last but not least, although the case $p = n$ is not covered by the mathematical treatment, it is also dealt with successfully in practice by the new nonlinear shrinkage estimator.

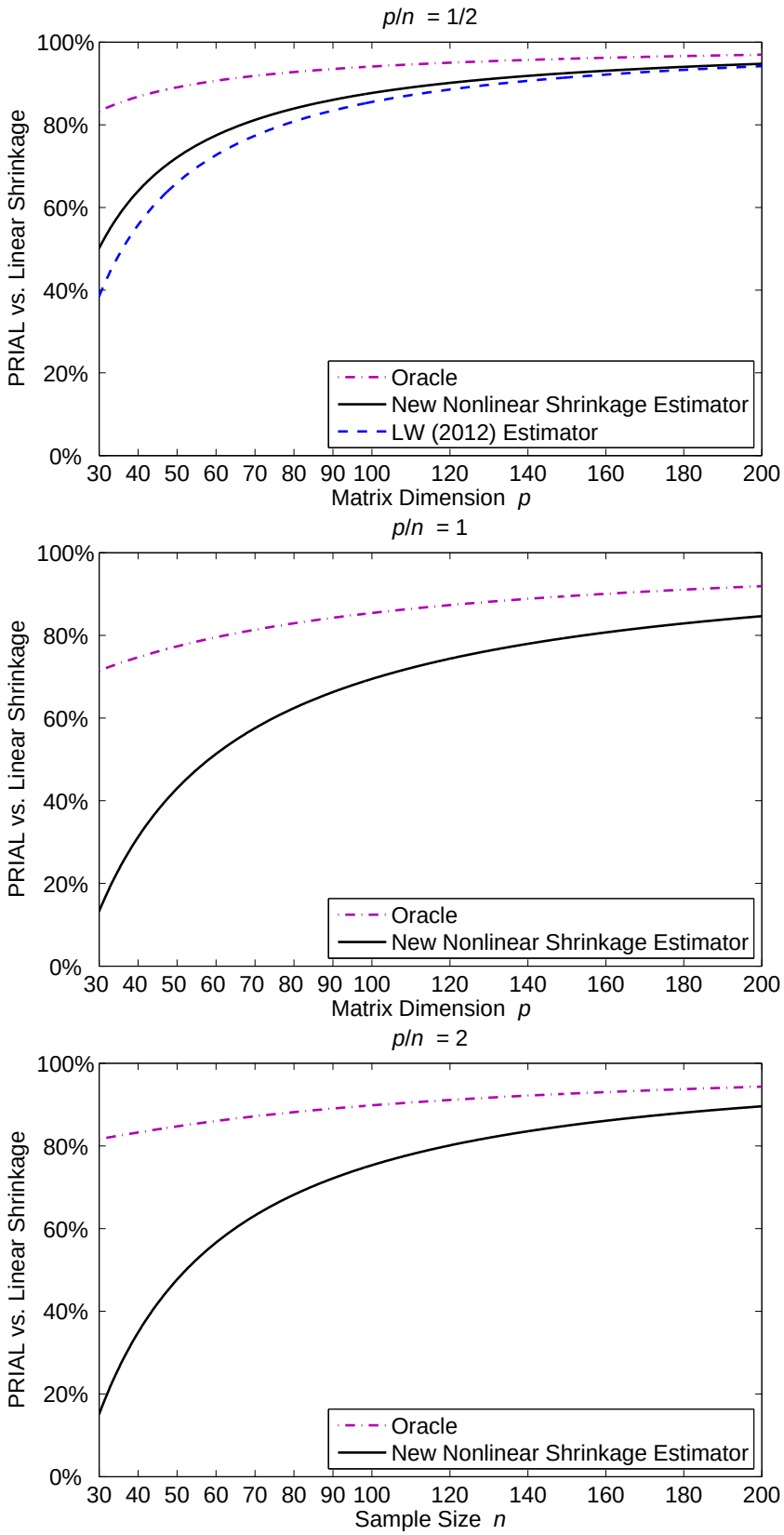

p/n = 1/2
PRIAL vs. Linear Shrinkage
100%
80%
60%
40%
20%
0%
30 40 50 60 70 80 90 100 120 140 160 180 200
Matrix Dimension p
Oracle
New Nonlinear Shrinkage Estimator
LW (2012) Estimator
p/n = 1
PRIAL vs. Linear Shrinkage
100%
80%
60%
40%
20%
0%
30 40 50 60 70 80 90 100 120 140 160 180 200
Matrix Dimension p
Oracle
New Nonlinear Shrinkage Estimator
p/n = 2
PRIAL vs. Linear Shrinkage
100%
80%
60%
40%
20%
0%
30 40 50 60 70 80 90 100 120 140 160 180 200
Sample Size n
Oracle
New Nonlinear Shrinkage Estimator

Figure 6: Percentage Improvement in Average Loss (PRIAL) according to the Frobenius norm of nonlinear versus linear shrinkage estimation of the covariance matrix.

### 5.3 Principal Component Analysis

Recall that in Section 4 on principal component analysis we dropped the first subscript $n$ always, and so the same will be done in this section.

In our design, the $i$th population eigenvalue is equal to $\tau_i = H^{-1}((i-0.5)/p)$ $(i = 1, \ldots, p)$, where $H$ is given by the distribution of $1 + 10W$, and $W \sim \text{Beta}(1, 10)$ The distribution of the random variates comprising the $n \times p$ data matrix $X_n$ is Gaussian. We consider the two cases $(n = 200, p = 100)$ and $(n = 100, p = 200)$, so the concentration is $p/n = 0.5$ or $p/n = 2$.

Let $y \in \mathbb{R}^p$ be a random vector with covariance matrix $\Sigma$, drawn independently from the sample covariance matrix $S$. The out-of-sample variance of the $i$th (sample) principal component, $u_i'y$, is given by $d_i^* := u_i'\Sigma u_i$; see (3.4). By our convention, the $d_i^*$ are sorted in increasing order.

We consider the following estimators of $d_i^*$.

- **Sample:** The estimator of $d_i^*$ is the $i$th sample eigenvalue, $\lambda_i$.
- **Population:** The estimator of $d_i^*$ is the $i$th population eigenvalue, $\tau_i$; this estimator is not feasible but is included for educational purposes nevertheless.
- **LW:** The estimator of $d_i^*$ is the nonlinear shrinkage quantity $\widehat{d}_i$ as given in equation (3.8) in the case $p < n$ and in equation (3.11) in the case $p > n$.

Let $\widetilde{d}_i$ be a generic estimator of $d_i^*$. First, we are plotting

$$\widetilde{f}_k := \frac{\sum_{j=1}^k \widetilde{d}_{p-j+1}}{\sum_{m=1}^p \widetilde{d}_m}$$

as a function of $k$, averaged over 1,000 Monte Carlo simulations. The quantity $\widetilde{f}_k$ serves as an estimator of $f_k$, the fraction of the total variation in $y$ that is explained by the $k$ largest principal components:

$$f_k := \frac{\sum_{j=1}^k d_{p-j+1}^*}{\sum_{m=1}^p d_m^*}$$

The results are displayed in Figure 7 (one panel for each case $p/n = 0.5, 2.$) The upward bias of Sample is apparent, while LW is very close to the Truth. Moreover, Population is also upward biased (though not as much as Sample): the important message is that even if the population eigenvalues were known, they should not be used to judge the variances of the (sample) principal components. As expected, the differences between Sample and LW increase with the concentration $p/n$; the same is true for the differences between Population and LW.

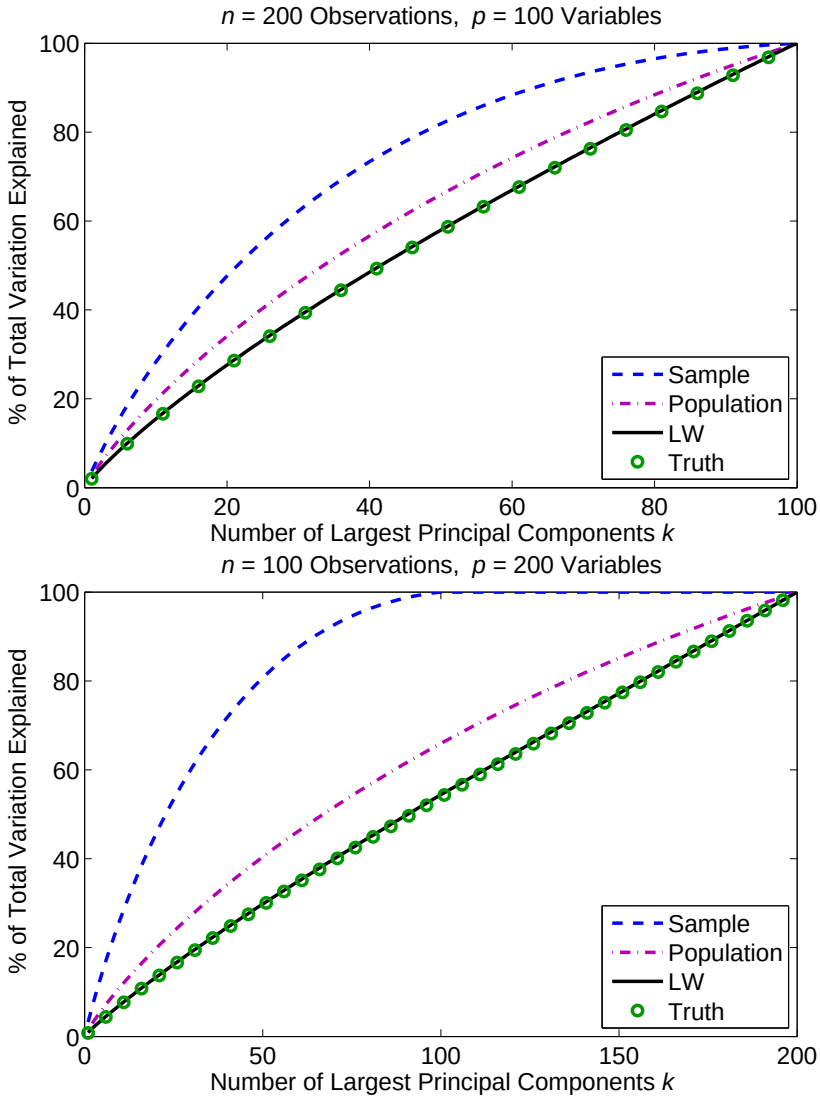



Figure 7: Comparison between different estimators of the percentage of total variation explained by the top principal components.

Figure 7 shows how close the estimator $\widetilde{f}_k$ is to the truth $f_k$ on average. But it does not necessarily answer how close the cumulative-percentage-of-total-variation rule based on $\widetilde{f}_k$ is to the rule based on $f_k$. For a given percentage $(q \times 100)\%$, with $q \in (0, 1)$, let

$$k(q) := \min\{k : f_k \geq q\} \qquad \text{and} \qquad \widetilde{k}(q) := \min\{k : \widetilde{f}_k \geq q\} \ .$$

In words, $k(q)$ is the (smallest) number of the largest principal components that must be retained to explain $(q \times 100)\%$ of the total variation and $\widetilde{k}(q)$ is an estimator of this quantity.

We are then also interested in the Root Mean Squared Error (RMSE) of $\widetilde{k}(q)$, defined as

$$\text{RMSE} := \sqrt{\mathbb{E}\left[\left(\widetilde{k}(q) - k(q)\right)^2\right]} \ ,$$

for the values of $q$ most commonly used in practice, namely $q = 0.7, 0.8, 0.9$. We compute empirical RMSEs across 1,000 Monte Carlo simulations. The results are presented in Table 4.

It can be seen that in each scenario, Sample has the largest RMSE and LW has the smallest RMSE; in particular LW is highly accurate not only relative to Sample but also in an absolute sense. (The quantity $k(q)$ is a random variable, since it depends on the sample eigenvalues $u_i$; but the general magnitude of $k(q)$ for the various scenarios can be judged from Figure 7)

| $q$ | Sample | Population | LW | | $q$ | Sample | Population | LW |
|---|---|---|---|---|---|---|---|---|
| $n = 200, p = 100$ | | | | | $n = 100, p = 200$ | | | |
| 70% | 26.9 | 9.1 | 0.8 | | 70% | 96.5 | 25.4 | 1.4 |
| 80% | 27.9 | 8.0 | 0.8 | | 80% | 107.3 | 21.0 | 0.9 |
| 90% | 24.4 | 5.0 | 0.6 | | 90% | 114.0 | 13.0 | 0.7 |

Table 4: Empirical Root Mean Squared Error (RMSE) of various estimates, $\widetilde{k}(q)$, of the number of largest principal components that must be retained, $k(q)$, to explain $(q \times 100)\%$ of the total variation. Based on 1,000 Monte Carlo simulations.

# 6 Empirical Application

As an empirical application, we study principal component analysis (PCA) in the context of stock return data. Principal components of a return vector of a cross section of $p$ stocks are used for risk analysis and portfolio selection by finance practitioners; for example, see Roll and Ross (1980) and Connor and Korajczyk (1993).

We use the $p = 30, 60, 240, 480$ largest stocks, as measured by their market value at the beginning of 2011, that have a complete return history from January 2002 until December 2011. As is customary in many financial applications, such as portfolio selection, we use monthly data. Consequently, the sample size for the ten-year history is $n = 120$ and the concentration is $p/n = 0.25, 0.5, 2, 4$.

It is of crucial interest how much of the total variation in the $p$-dimensional return vector is explained by the $k$ largest principal components. We compare the approach based on the sample covariance matrix, defined in equation (4.2) and denoted by Sample, to that of nonlinear shrinkage, defined in equation 4.3 and denoted by LW. The results are displayed in Figure 8. It can be seen that Sample is overly optimistic compared to LW and, as expected, the differences between the two methods increase with the concentration $p/n$.

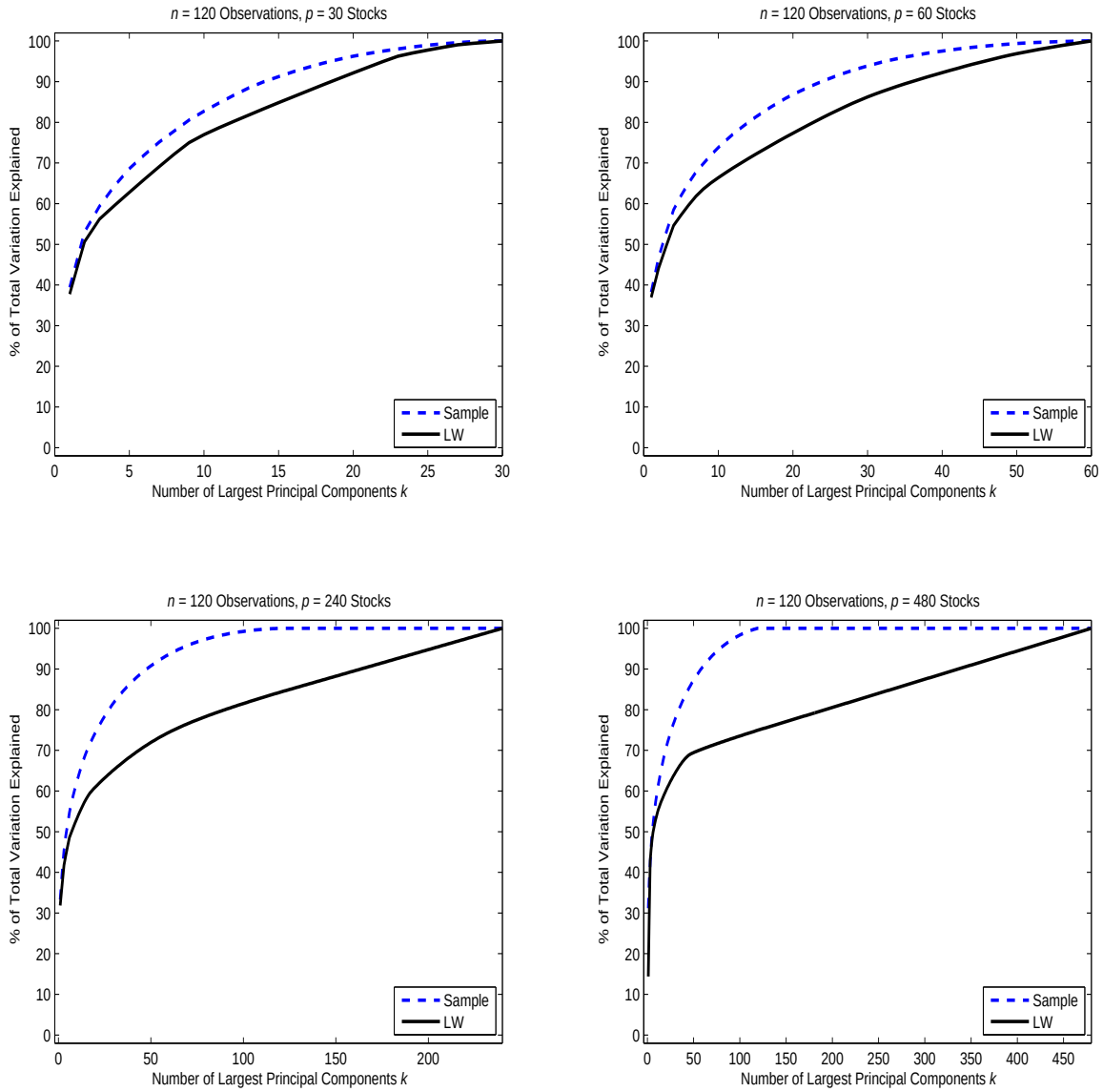



Figure 8: Percentage of total variation explained by the $k$ largest principal components of stock returns: estimates based on the sample covariance matrix (Sample) compared to those based on nonlinear shrinkage (LW).

In addition to the visual analysis, the differences can also be presented via the cumulative-percentage-of-total-variation rule to decide how many of the largest principal components to retain; see Section 4. The results are presented in Table 5. It can be seen again that Sample is overly optimistic compared to LW and retains much fewer principal components. Again, as expected, the differences between the two methods increase with the concentration $p/n$.

| $f_{Target}$ | LW | Sample | | $f_{Target}$ | LW | Sample |
|---|---|---|---|---|---|---|
| $n = 120, p = 30$ | | | | $n = 120, p = 60$ | | |
| 70% | 8 | 6 | | 70% | 14 | 9 |
| 80% | 12 | 9 | | 80% | 23 | 15 |
| 90% | 19 | 15 | | 90% | 36 | 24 |
| $n = 120, p = 240$ | | | | $n = 120, p = 480$ | | |
| 70% | 44 | 16 | | 70% | 56 | 20 |
| 80% | 91 | 28 | | 80% | 193 | 34 |
| 90% | 164 | 48 | | 90% | 337 | 59 |

Table 5: Number of $k$ largest principal components to retain according to the cumulative-percentage-of-total-variation rule. The rule is based either on the sample covariance matrix (Sample) or on nonlinear shrinkage (LW).

# 7 Conclusion

The analysis of large-dimensional data sets is becoming more and more common. For many statistical problems, the classic textbook methods no longer work well in such settings. Two cases in point are covariance matrix estimation and principal component analysis, both cornerstones of multivariate analysis.

The classic estimator of the covariance matrix is the sample covariance matrix. It is unbiased and the maximum-likelihood estimator under normality. But when the dimension is not small compared to the sample size, the sample covariance matrix contains too much estimation error and is ill-conditioned; when the dimension is larger than the sample size, it is not even invertible anymore.

The variances of the principal components (which are obtained from the sample eigenvectors) are no longer accurately estimated by the sample eigenvalues. In particular, the sample eigenvalues overestimate the variances of the 'large' principal components. As a result the common rules in determining how many principal components to retain generally select too few of them.

In the absence of strong structural assumptions on the true covariance matrix, such as sparseness or a factor model, a common remedy for both statistical problems is *nonlinear shrinkage* of the sample eigenvalues. The optimal shrinkage formula delivers an estimator of the covariance matrix that is finite-sample optimal with respect to the Frobenius norm in the class of rotation-equivariant estimators. The *same* shrinkage formula also gives the variances of the (sample) principal components. It is noteworthy that the optimal shrinkage formula is different from the population eigenvalues: even if they were available, one should not use them for the ends of covariance matrix estimation and PCA.

Unsurprisingly, the optimal nonlinear shrinkage formula is not available, since it depends on population quantities. But an asymptotic counterpart, denoted *oracle shrinkage*, can be estimated consistently. In this way, *bona fide* nonlinear shrinkage estimation of covariance matrices and improved PCA result.

The key to the consistent estimation of the oracle shrinkage is the consistent estimation of the population eigenvalues. This problem is challenging and interesting in its own right, solving a host of additional statistical problems. Our proposal to this end is the first one that does not make strong assumptions on the distribution of the population eigenvalues, has proven consistency, and also works well in practice.

Extensive Monte Carlo simulations have established that our methods have desirable finite-sample properties and outperform methods that have been previously suggested in the literature.

## References


Amini, A. A. (2011). High-dimensional principal component analysis. Technical Report UCB/EECS-2011-104, Department of Electrical Engineering and Computer Sciences, University of California at Berkeley.

Anatolyev, S. (2012). Inference in regression models with many regressors. *Journal of Econometrics*, 170(2):368–382.

Bai, Z. D. and Silverstein, J. W. (1998). No eigenvalues outside the suppport of the limiting spectral distribution of large-dimensional random matrices. *Annals of Probability*, 26(1):316–345.

Bai, Z. D. and Silverstein, J. W. (1999). Exact separation of eigenvalues of large-dimensional sample covariance matrices. *Annals of Probability*, 27(3):1536–1555.

Bai, Z. D. and Silverstein, J. W. (2010). *Spectral Analysis of Large-Dimensional Random Matrices*. Springer, New York, second edition.

Bickel, P. J. and Freedman, D. A. (1981). Asymptotic theory for the bootstrap. *Annals of Statistics*, 9(6):1196–1217.

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Annals of Statistics*, 36(1):199–227.

Connor, G. and Korajczyk, R. A. (1993). A test for the number of factors in an approximate factor model. *Journal of Finance*, 48:1263–1291.

Demetrescu, M. and Hanck, C. (2012). A simple nonstationary-volatility robust panel unit root test. *Economics Letters*, 117(1):10–13.

El Karoui, N. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Annals of Statistics*, 36(6):2757–2790.

Fan, J., Fan, Y., and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197.

Gill, P. E., Murray, W., and Saunders, M. A. (2002). SNOPT: An SQP algorithm for large-scale constrained optimization. *SIAM Journal on Optimization*, 12(4):979–1006.

Guo, S.-M., He, J., Monnier, N., Sun, G., Wohland, T., and Bathe, M. (2012). Bayesian approach to the analysis of fluorescence correlation spectroscopy data II: Application to simulated and in vitro data. *Analytical Chemistry*, 84(9):3880–3888.

Haufe, S., Treder, M., Gugler, M., Sagebaum, M., Curio, G., and Blankertz, B. (2011). EEG potentials predict upcoming emergency brakings during simulated driving. *Journal of Neural Engineering*, 8(5).

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, 498–520.

Huang, T.-K. and Schneider, J. (2011). Learning auto-regressive models from sequence and non-sequence data. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 24*, pages 1548–1556. The MIT Press, Cambridge.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer, New York, second edition.

Khan, M. (2008). Are accruals mispriced? Evidence from tests of an intertemporal capital asset pricing model. *Journal of Accounting and Economics*, 45(1):55–77.

Lawley, D. N. (1956). A general method for approximating to the distribution of likelihood ratio criteria. *Biometrika*, 43(3/4):295–303.

Ledoit, O. and Péché, S. (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 150(1–2):233–264.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411.

Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Annals of Statistics*, 40(2):1024–1060.

Lin, J. and Bentler, P. (2012). A third moment adjusted test statistic for small sample factor analysis. *Multivariate Behavioral Research*, 47(3):448–462.

Lin, J.-A., Zhu, H. b., Knickmeyer, R., Styner, M., Gilmore, J., and Ibrahim, J. (2012). Projection regression models for multivariate imaging phenotype. *Genetic Epidemiology*, 36(6):631–641.

Marčenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Sbornik: Mathematics*, 1(4):457–483.

Markon, K. (2010). Modeling psychopathology structure: A symptom-level analysis of axis I and II disorders. *Psychological Medicine*, 40(2):273–288.

Mestre, X. (2008). Improved estimation of eigenvalues and eigenvectors of covariance matrices using their sample estimates. *IEEE Transactions on Information Theory*, 54(11):5113–5129.

Nguyen, L., Rheinschmitt, R., Wild, T., and Brink, S. (2011). Limits of channel estimation and signal combining for multipoint cellular radio (comp). In *Wireless Communication Systems (ISWCS), 2011 8th International Symposium on*, pages 176–180. IEEE.

Pearson, K. (1901). On line and planes of closest fit to systems of points in space. *Philosophical Magazine (Series 6)*, 2(11):559–572.

Pedro Duarte Silva, A. (2011). Two-group classification with high-dimensional correlated data: A factor model approach. *Computational Statistics and Data Analysis*, 55(11):2975–2990.

Perlman, M. D. (2007). *STAT 542: Multivariate Statistical Analysis*. University of Washington (On-Line Class Notes), Seattle, Washington.

Pirkl, R., Remley, K., and Patan, C. (2012). Reverberation chamber measurement correlation. *IEEE Transactions on Electromagnetic Compatibility*, 54(3):533–545.

Pyeon, D., Newton, M., Lambert, P., Den Boon, J., Sengupta, S., Marsit, C., Woodworth, C., Connor, J., Haugen, T., Smith, E., Kelsey, K., Turek, L., and Ahlquist, P. (2007). Fundamental differences in cell cycle deregulation in human papillomavirus-positive and human papillomavirus-negative head/neck and cervical cancers. *Cancer Research*, 67(10):4605–4619.

Rajaratnam, B., Massam, H., and Carvalho, C. M. (2008). Flexible covariance estimation in graphical Gaussian models. *Annals of Statistics*, 36(6):2818–2849.

Ribes, A., Azas, J.-M., and Planton, S. (2009). Adaptation of the optimal fingerprint method for climate change detection using a well-conditioned covariance matrix estimate. *Climate Dynamics*, 33(5):707–722.

Roll, R. and Ross, S. A. (1980). An empirical investigation of the arbitrage pricing theory. *Journal of Finance*, 35:1073–1103.

Sætrom, J., Hove, J., Skjervheim, J.-A., and Vabø, J. (2012). Improved uncertainty quantification in the ensemble Kalman filter using statistical model-selection techniques. *SPE Journal*, 17(1):152–162.

Silverstein, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. *Journal of Multivariate Analysis*, 55:331–339.

Silverstein, J. W. and Bai, Z. D. (1995). On the empirical distribution of eigenvalues of a class of large-dimensional random matrices. *Journal of Multivariate Analysis*, 54:175–192.

Silverstein, J. W. and Choi, S. I. (1995). Analysis of the limiting spectral distribution of large-dimensional random matrices. *Journal of Multivariate Analysis*, 54:295–309.

Stein, C. (1975). Estimation of a covariance matrix. Rietz lecture, 39th Annual Meeting IMS. Atlanta, Georgia.

Stein, C. (1986). Lectures on the theory of estimation of many parameters. *Journal of Mathematical Sciences*, 34(1):1373–1403.

Tsagaris, T., Jasra, A., and Adams, N. (2012). Robust and adaptive algorithms for online portfolio selection. *Quantitative Finance*, 12(11):1651–1662.

Varoquaux, G., Gramfort, A., Poline, J.-B., and Thirion, B. (2010). Brain covariance selection: better individual functional connectivity models using population prior. In Lafferty, J., Williams, C. K. I., Shawe-Taylor, J., Zemel, R., and Culotta, A., editors, *Advances in Neural Information Processing Systems 23*, pages 2334–2342. The MIT Press, Cambridge.

Wei, Z., Huang, J., and Hui, Y. (2011). Adaptive-beamforming-based multiple targets signal separation. In *Signal Processing, Communications and Computing (ICSPCC), 2011 IEEE International Conference on*, pages 1–4. IEEE.

Yao, J., Kammoun, A., and Najim, J. (2012). Estimation of the covariance matrix of large dimensional data. [Online]. Available at `http://arxiv.org/abs/1201.4672`.

# A Mathematical Proofs

**Lemma A.1.** *Let $\{G_n\}$ and $G$ be c.d.f.'s on the real line and assume that there exists a compact interval that contains the support of $G$ as well as the support of $G_n$ for all $n$ large enough. For $0 < \alpha < 1$, let $G_n^{-1}(\alpha)$ denote an $\alpha$ quantile of $G_n$ and let $G^{-1}(\alpha)$ denote an $\alpha$ quantile of $G$. Let $\{K_n\}$ be a sequence of integers with $K_n \to \infty$; further, let $t_{n,k} := (k - 0.5)/K_n$ $(k = 1, \ldots, K_n)$.*

*Then $G_n \Rightarrow G$ if and only if*

$$\frac{1}{K_n} \sum_{k=1}^{K_n} \left[G_n^{-1}(t_{n,k}) - G^{-1}(t_{n,k})\right]^2 \to 0 \ . \tag{A.1}$$

Proof. First, since there exists a compact interval that contains the support of $G$ as well as the support of $G_n$ for all $n$ large enough, weak convergence of $G_n$ to $G$ implies that also the second moment of $G_n$ converges to the second moment of $G$.

Second, we claim that, under the given set of assumptions, the convergence (A.1) is equivalent to the convergence

$$\int_0^1 \left[G_n^{-1}(t) - G^{-1}(t)\right]^2 dt \to 0 \ . \tag{A.2}$$

If this claim is true, the proof of the lemma then follows from Lemmas 8.2 and 8.3(a) of Bickel and Freedman (1981). We are, therefore, left to show the claim that the convergence (A.1) is equivalent to the convergence (A.2).

We begin by showing that the convergence (A.1) implies the convergence (A.2). One can make the following decomposition:

$$\begin{aligned}\int_0^1 \left[G_n^{-1}(t) - G^{-1}(t)\right]^2 dt = & \int_0^{t_{n,1}} \left[G_n^{-1}(t) - G^{-1}(t)\right]^2 dt \\ & + \int_{t_{n,1}}^{t_{n,K_n}} \left[G_n^{-1}(t) - G^{-1}(t)\right]^2 dt \\ & + \int_{t_{n,K_n}}^1 \left[G_n^{-1}(t) - G^{-1}(t)\right]^2 dt \ .\end{aligned}$$

Let $C$ denote the length of the compact interval that contains the support of $G$ and $G_n$ (for all $n$ large enough). Also, note that $t_{n,1} - 0 = 1 - t_{n,K_n} = 0.5/K_n$. Then, for all $n$ large enough, we can use the trivial bound

$$\int_0^{t_{n,1}} \left[G_n^{-1}(t) - G^{-1}(t)\right]^2 dt + \int_{t_{n,K_n}}^1 \left[G_n^{-1}(t) - G^{-1}(t)\right]^2 dt \le \frac{0.5}{K_n} C^2 + \frac{0.5}{K_n} C^2 = \frac{C^2}{K_n} \ .$$

Combining this bound with the previous decomposition results in

$$\int_0^1 \left[G_n^{-1}(t) - G^{-1}(t)\right]^2 dt \le \frac{C^2}{K_n} + \int_{t_{n,1}}^{t_{n,K_n}} \left[G_n^{-1}(t) - G^{-1}(t)\right]^2 dt \ ,$$

and we are left to show that

$$\int_{t_{n,1}}^{t_{n,K_n}} \left[G_n^{-1}(t) - G^{-1}(t)\right]^2 dt \to 0 \ .$$

For any $k = 1, \ldots, K_n - 1$, noting that $t_{n,k+1} - t_{n,k} = 1/K_n$,

$$\int_{t_{n,k}}^{t_{n,k+1}} \left[G_n^{-1}(t) - G^{-1}(t)\right]^2 dt \leq \frac{1}{K_n} \sup_{t_{n,k} \leq t \leq t_{n,k+1}} \left[G_n^{-1}(t) - G^{-1}(t)\right]^2$$
$$\leq \frac{\left[G_n^{-1}(t_{n,k+1}) - G^{-1}(t_{n,k})\right]^2 + \left[G_n^{-1}(t_{n,k}) - G^{-1}(t_{n,k+1})\right]^2}{K_n} ,$$

where the last inequality follows from the fact that both $G_n^{-1}$ and $G^{-1}$ are (weakly) increasing functions. As a result,

$$\int_{t_{n,1}}^{t_{n,K_n}} \left[G_n^{-1}(t) - G^{-1}(t)\right]^2 dt = \sum_{k=1}^{K_n-1} \int_{t_{n,k}}^{t_{n,k+1}} \left[G_n^{-1}(t) - G^{-1}(t)\right]^2$$
$$\leq \quad \frac{1}{K_n} \sum_{k=1}^{K_n-1} \left[G_n^{-1}(t_{n,k+1}) - G^{-1}(t_{n,k})\right]^2 \tag{A.3}$$
$$+ \frac{1}{K_n} \sum_{k=1}^{K_n-1} \left[G_n^{-1}(t_{n,k}) - G^{-1}(t_{n,k+1})\right]^2 , \tag{A.4}$$

and we are left to show that both terms (A.3) and (A.4) converge to zero.

The term (A.3) can be written as

$$\text{(A.3)} = \frac{1}{K_n} \sum_{k=1}^{K_n-1} \left[G_n^{-1}(t_{n,k}) - G^{-1}(t_{n,k}) + a_{n,k}\right]^2 ,$$

with $a_{n,k} \coloneqq G_n^{-1}(t_{n,k+1}) - G_n^{-1}(t_{n,k})$; note that $\sum_{k=1}^{K_n-1} |a_{n,k}| \leq C$.

Next, write

$$\frac{1}{K_n} \sum_{k=1}^{K_n-1} \left[G_n^{-1}(t_{n,k}) - G^{-1}(t_{n,k}) + a_{n,k}\right]^2 = \quad \frac{1}{K_n} \sum_{k=1}^{K_n-1} \left[G_n^{-1}(t_{n,k}) - G^{-1}(t_{n,k})\right]^2 \tag{A.5}$$
$$+ \frac{2}{K_n} \sum_{k=1}^{K_n-1} \left[G_n^{-1}(t_{n,k}) - G^{-1}(t_{n,k})\right] \cdot a_{n,k} \tag{A.6}$$
$$+ \frac{1}{K_n} \sum_{k=1}^{K_n-1} a_{n,k}^2 . \tag{A.7}$$

The term on the right-hand side (A.5) converges to zero by assumption. The term (A.7) converges to zero because

$$\sum_{k=1}^{K_n-1} a_{n,k}^2 \leq \Big( \sum_{k=1}^{K_n-1} |a_{n,k}| \Big)^2 \leq C^2 .$$

Since both the term on the right-hand side of (A.5) and the term (A.7) converge to zero, the term (A.6) converges to zero as well by the Cauchy-Schwarz inequality. Consequently, the term (A.3) converges to zero.

By a completely analogous argument, the term (A.4) converges to zero too.

We have thus established that the convergence (A.1) implies the convergence (A.2). By a similar argument, one can establish the reverse fact that the convergence (A.2) implies the convergence (A.1). ■

It is useful to discuss Lemma A.1 a bit further. For two c.d.f.'s $G_1$ and $G_2$ on the real line, define

$$||G_1 - G_2||_p := \sqrt{\frac{1}{p}\sum_{i=1}^{p}\left[G_1^{-1}((i-0.5)/p) - G_2^{-1}((i-0.5)/p)\right]^2} \ . \tag{A.8}$$

Two results are noted.

First, the left-hand expression of equation (A.1) can be written as

$$||G_n - G||_p^2$$

when $p = K_n$. Since $p \to \infty$, Lemma A.1 states that, under the given set of assumptions,

$$G_n \Rightarrow G \quad \text{ if and only } \quad ||G_n - G||_p^2 \to 0 \ . \tag{A.9}$$

Second, a triangular inequality holds in the sense that for three c.d.f.'s $G_1, G_2$, and $G_3$ on the real line,

$$||G_1 - G_2||_p \le ||G_1 - G_3||_p + ||G_2 - G_3||_p \ . \tag{A.10}$$

This second fact follows since, for example, $\sqrt{p} \cdot ||G_1 - G_2||_p$ is the Euclidian distance between the two vectors $(G_1^{-1}(0.5/p), \ldots, G_1^{-1}((p-0.5)/p))'$ and $(G_2^{-1}(0.5/p), \ldots, G_2^{-1}((p-0.5)/p))'$.

These two results are summarized in the following corollary.

**Corollary A.1.**

*(i) Let $\{G_n\}$ and $G$ be c.d.f.'s on the real line and assume that there exists a compact interval that contains the support of $G$ as well as the support of $G_n$ for all $n$ large enough. For $0 < \alpha < 1$, let $G_n^{-1}(\alpha)$ denote an $\alpha$ quantile of $G_n$ and let $G^{-1}(\alpha)$ denote an $\alpha$ quantile of $G$. Also assume that $p \to \infty$.*

*Then $G_n \Rightarrow G$ if and only if*

$$||G_n - G||_p^2 \to 0 \ ,$$

*where $|| \cdot ||_p$ is defined as in* (A.8).

*(ii) Let $G_1$, $G_2$, and $G_3$ be c.d.f.'s on the real line. Then*

$$||G_1 - G_2||_p \le ||G_1 - G_3||_p + ||G_2 - G_3||_p \ .$$

Proof of Theorem 2.1. As shown by Silverstein (1995), $F_n \Rightarrow F$ almost surely. Therefore, by Corollary A.1(i),

$$\frac{1}{p}\sum_{i=1}^{p}[\lambda_{n,i} - F^{-1}((i-0.5)/p)]^2 \xrightarrow{\text{a.s.}} 0 \ , \tag{A.11}$$

recalling that $\lambda_{n,i}$ is a $(i-0.5)/p$ quantile of $F_n$; see Remark 2.1. The additional fact that

$$\frac{1}{p}\sum_{i=1}^{p}\left[q_{n,p}^{i}(\boldsymbol{\tau}_n) - F^{-1}((i-0.5)/p)\right]^2 \xrightarrow{\text{a.s.}} 0 \tag{A.12}$$

follows from the Marčenko-Pastur equation (2.5), Lemma A.2 of Ledoit and Wolf (2012), Assumption (A3), the definition of $q_{n,p}^{i}(\boldsymbol{\tau}_n)$, and Corollary A.1(i) again. The convergences (A.11) and (A.12) together with the triangular inequality for the Euclidian distance in $\mathbb{R}^p$ then imply that

$$\frac{1}{p}\sum_{i=1}^{p}\left[q_{n,p}^{i}(\boldsymbol{\tau}_n) - \lambda_{n,i}\right]^2 \xrightarrow{\text{a.s.}} 0 \ ,$$

which is the statement to be proven. ■

Proof of Theorem 2.2. For any probability measure $\widetilde{H}$ on the nonnegative real line and for any $\widetilde{c} > 0$, let $F_{\widetilde{H},\widetilde{c}}$ denote the c.d.f. on the real line induced by the corresponding solution of the Marčenko-Pastur equation (2.5). More specifically, for each $z \in \mathbb{C}^+$, $m_{F_{\widetilde{H},\widetilde{c}}}(z)$ is the unique solution for $m \in \mathbb{C}^+$ to the equation

$$m = \int_{-\infty}^{+\infty} \frac{1}{\tau\left[1-\widetilde{c}-\widetilde{c}\,z\,m\right]-z}\, d\widetilde{H}(\tau) \ .$$

In this notation, $F = F_{H,c}$.

Recall that $F_n$ denotes the empirical c.d.f. of the sample eigenvalues $\boldsymbol{\lambda}_n$. Furthermore, for $\mathbf{t} := (t_1, \ldots, t_p)' \in [0,\infty)^p$, denote by $\widetilde{H}_{\mathbf{t}}$ the probability distribution that places mass $1/p$ at each of the $t_i$ $(i = 1, \ldots, p)$. The objective function in equation (2.19) can then be re-expressed as

$$||F_{\widetilde{H}_{\mathbf{t}},\widehat{c}_n} - F_n||_p^2 \ ,$$

where $||\cdot||_p$ is defined as in (A.8). Note here that $F_{\widetilde{H}_{\mathbf{t}},\widehat{c}_n}$ is nothing else than $F_{n,p}^{\mathbf{t}}$ of equation (2.15); but for the purposes of this proof, the notation $F_{\widetilde{H}_{\mathbf{t}},\widehat{c}_n}$ is more convenient.

Consider the following infeasible estimator of the limiting spectral distribution $H$:

$$\overline{H}_n := \underset{\widetilde{H}}{\operatorname{argmin}}\, ||F_{\widetilde{H},\widehat{c}_n} - F_n||_p^2 \ , \tag{A.13}$$

where the minimization is over *all* probability measures $\widetilde{H}$ on the real line; the estimator $\overline{H}_n$ is infeasible, since one cannot minimize over all probability measures on the real line in practice. By definition,

$$||F_{\overline{H}_n,\widehat{c}_n} - F_n||_p \leq ||F_{H,\widehat{c}_n} - F_n||_p \ . \tag{A.14}$$

Therefore,

$$\begin{aligned}
||F_{\overline{H}_n,\widehat{c}_n} - F||_p &\leq ||F_{\overline{H}_n,\widehat{c}_n} - F_n||_p + ||F_n - F||_p \quad \text{(by Corollary A.1(ii))} \\
&\leq ||F_{H,\widehat{c}_n} - F_n||_p + ||F_n - F||_p \quad \text{(by (A.14)} \\
&\leq ||F_{H,\widehat{c}_n} - F_{H,c}||_p + ||F_{H,c} - F_n||_p + ||F_n - F||_p \quad \text{(by Corollary A.1(ii))} \\
&= ||F_{H,\widehat{c}_n} - F||_p + 2\,||F_n - F||_p \quad \text{(since } F_{H,c} = F\text{)} \\
&=: A + B \ .
\end{aligned}$$

In the case $c < 1$, combining Corollary A.1(i) with Lemma A.2 of Ledoit and Wolf (2012) shows that $A \to 0$ almost surely. In the case $c > 1$, one can also show that $A \to 0$ almost surely: Lemma A.2 of Ledoit and Wolf (2012) implies that $\underline{F}_{H,\widehat{c}_n} \Rightarrow \underline{F}$ almost surely; then use equation (2.7) together with the fact that $\widehat{c}_n \to c$ to deduce that also $F_{H,\widehat{c}_n} \Rightarrow F$ almost surely; finally apply Corollary A.1(i). Combining Corollary A.1(i) with the fact that $F_n \Rightarrow F$ almost surely (Silverstein, 1995) shows that $B \to 0$ almost surely in addition to $A \to 0$ almost surely. Therefore, $||F_{\overline{H}_n,\widehat{c}_n} - F||_p \to 0$ almost surely. Using Corollary A.1(i) again shows that $F_{\overline{H}_n,\widehat{c}_n} \Rightarrow F$ almost surely.

A feasible estimator of $H$ is given by

$$\widehat{H}_n \coloneqq \underset{\widetilde{H}_{\mathbf{t}} \in \mathcal{P}_n}{\operatorname{argmin}} ||F_{\widetilde{H}_{\mathbf{t}},\widehat{c}_n} - F_n||_p^2$$

instead of by (A.13), where the subset $\mathcal{P}_n$ denotes the set of probability measures that are equal-weighted mixtures of $p$ point masses on the nonnegative real line:

$$\mathcal{P}_n \coloneqq \left\{ \widetilde{H}_{\mathbf{t}} : \widetilde{H}_{\mathbf{t}}(x) \coloneqq \frac{1}{p} \sum_{i=1}^{p} \mathbb{1}_{\{x \geq t_i\}} \text{ , where } \mathbf{t} \coloneqq (t_1, \ldots, t_p)' \in [0, \infty)^p \right\} .$$

The fact that the minimization over a finite but dense family of probability measures, instead of all probability measures on the nonnegative real line, does not affect the strong consistency of the estimator of $F$ follows by arguments similar to those used in the proof of Corollary 5.1(i) of Ledoit and Wolf (2012). Therefore, it also holds that $F_{\widehat{H}_n,\widehat{c}_n} \Rightarrow F$ almost surely.

Having established that $F_{\widehat{H}_n,\widehat{c}_n} \Rightarrow F$ almost surely, it follows that also $\widehat{H}_n \Rightarrow H$ almost surely; see the proof of Theorem 5.1(ii) of Ledoit and Wolf (2012). Since $\widehat{H}_n$ is recognized as the empirical distribution (function) of the $\widehat{\tau}_{n,i}$ $(i = 1, \ldots, p)$, $\widehat{\tau}_{n,i}$ is a $(i - 0.5)/p$ quantile of $\widehat{H}_n$; see Remark 2.1. Therefore, it follows from Corollary A.1(i) that

$$\frac{1}{p} \sum_{i=1}^{p} [\widehat{\tau}_{n,i} - H^{-1}((i - 0.5)/p)]^2 \xrightarrow{\text{a.s.}} 0 \text{ ,} \tag{A.15}$$

The additional fact that

$$\frac{1}{p} \sum_{i=1}^{p} \left[\tau_{n,i} - H^{-1}((i - 0.5)/p)\right]^2 \xrightarrow{\text{a.s.}} 0 \tag{A.16}$$

follows directly from Assumption (A3) and Corollary A.1(i) again. The convergences (A.15) and (A.16) together with the triangular inequality for the Euclidian distance in $\mathbb{R}^p$ then imply that

$$\frac{1}{p} \sum_{i=1}^{p} [\widehat{\tau}_{n,i} - \tau_{n,i}]^2 \xrightarrow{\text{a.s.}} 0 \text{ ,}$$

which is the statement to be proven. ■

Proof of Theorem 3.1. The claim for the case $p < n$ follows immediately from Proposition 4.3(ii) of Ledoit and Wolf (2012).

To treat the case $p > n$, let $j$ denote the smallest integer for which $\lambda_i > 0$. Note that $(j-1)/p \to (c-1)/c$ almost surely by the results of Bai and Silverstein (1999); indeed, since the $\lambda_i$ are sorted in increasing order, $(j-1)/p$ is just the fraction of sample eigenvalues that are equal to zero.

Now restrict attention to the set of probability one on which $\widehat{\breve{m}_{\underline{F}}(0)} \to \breve{m}_{\underline{F}}(0)$, $\widehat{H}_n \Rightarrow H$, and $(j-1)/p \to (c-1)/c$. Adapting Proposition 4.3(i)(a) of Ledoit and Wolf (2012) to the continuous part of $F$, it can be shown that $\breve{m}_{\widehat{H}_n,\widehat{c}_n}(\lambda) \to \breve{m}_F(\lambda)$ uniformly in $\lambda \in \mathsf{Supp}(\underline{F})$, except for two arbitrarily small regions at the lower and upper end of $\mathsf{Supp}(\underline{F})$. We can write

$$
\begin{aligned}
||\widehat{S}_n - S_n^{or}||_F^2 &= \frac{j-1}{p}\left(\frac{1/\widehat{c}_n}{(1-1/\widehat{c}_n)\,\widehat{\breve{m}_{\underline{F}}(0)}} - \frac{1/c}{(1-1/c)\,\breve{m}_{\underline{F}}(0)}\right)^2 \\
&\quad + \frac{1}{p}\sum_{i=j}^{p}\left(\frac{\lambda_i}{\left|1-\widehat{c}_n-\widehat{c}_n\,\lambda_i\,\breve{m}_{F_{\widehat{H}_n,\widehat{c}_n}}(\lambda_i)\right|^2} - \frac{\lambda_i}{\left|1-c-c\,\lambda_i\,\breve{m}_F(\lambda_i)\right|^2}\right)^2 \\
&=: D_1 + D_2\ . \qquad \text{(A.17)}
\end{aligned}
$$

The fact that the summand $D_1$ converges to zero is obvious, keeping in mind that $c > 1$ and $\breve{m}_{\underline{F}}(0) > 0$. The fact that the summand $D_2$ converges to zero follows by arguments similar to those in the proof of Proposition 4.3(i)(b) of Ledoit and Wolf (2012).

We have thus shown that there exists a set of probability one on which $||\widehat{S}_n - S_n^{or}||_F \to 0$. ■

# B Justification of Remark 4.1

NOTATION.

- Let $y$ be a real $p$-dimensional random vector with covariance matrix $\Sigma$.
- Let $I_k$ denote the $k$-dimensional identity matrix, where $1 \leq k \leq p$.
- Let $W$ be a real nonrandom matrix of dimension $p \times k$ such that $W'W = I_k$.
- Let $w_i$ denote the $i$th column vector of $W$ $(i = 1, \ldots, k)$.

We start from the following two statements.

(1) If $\mathsf{Cov}[w_i'y, w_j'y] = 0$ for all $i \neq j$ then the variation attributable to the set of random variables $(w_1'y, \ldots, w_k'y)$ is $\sum_{i=1}^k \mathsf{Var}[w_i'y]$.

(2) If $R$ is a $k \times k$ rotation matrix, that is, $R'R = RR' = I_k$, and $\widetilde{w}_i$ is the $i$th column vector of the matrix $WR$, then the variation attributable to the rotated variables $(\widetilde{w}_1'y, \ldots, \widetilde{w}_k'y)$ is the same as the variation attributable to the original variables $(w_1'y, \ldots, w_k'y)$.

Together, Statements (1) and (2) imply that, even if $\mathsf{Cov}[w_i'y, w_j'y] \neq 0$ for $i \neq j$, the variation attributable to $(w_1'y, \ldots, w_k'y)$ is still $\sum_{i=1}^k \mathsf{Var}[w_i'y]$.

PROOF. Let us choose $R$ as a matrix of eigenvectors of $W'\Sigma W$. Then $(WR)'\Sigma(WR)$ is diagonal and $(WR)'(WR) = I_k$. Therefore, by Statement (1), the variation attributable to $(\widetilde{w}_1'y, \ldots, \widetilde{w}_k'y)$ is $\sum_{i=1}^k \mathsf{Var}[\widetilde{w}_i'y] = \mathsf{Tr}[(WR')\Sigma(WR)]$. By the properties of the trace operator, this is equal to $\mathsf{Tr}(W'\Sigma W) = \sum_{i=1}^k \mathsf{Var}[w_i'y]$. By Statement (2), it is the same as the variation attributable to $(w_1'y, \ldots, w_k'y)$. ■