Ledoit, Olivier; Wolf, Michael

Working Paper

Optimal estimation of a large-dimensional covariance matrix under Stein's loss

Working Paper, No. 122

Provided in Cooperation with:
Department of Economics, University of Zurich

Suggested Citation: Ledoit, Olivier; Wolf, Michael (2013) : Optimal estimation of a large-dimensional covariance matrix under Stein's loss, Working Paper, No. 122, University of Zurich, Department of Economics, Zurich,
https://doi.org/10.5167/uzh-78074

This Version is available at:
https://hdl.handle.net/10419/77592

University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 122

Optimal Estimation of a Large-Dimensional Covariance Matrix under Stein's Loss

Olivier Ledoit and Michael Wolf

May 2013

Optimal Estimation of a Large-Dimensional Covariance Matrix under Stein's Loss

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

May 2013

Abstract

This paper revisits the methodology of Stein (1975, 1986) for estimating a covariance matrix in the setting where the number of variables can be of the same magnitude as the sample size. Stein proposed to keep the eigenvectors of the sample covariance matrix but to shrink the eigenvalues. By minimizing an unbiased estimator of risk, Stein derived an 'optimal' shrinkage transformation. Unfortunately, the resulting estimator has two pitfalls: the shrinkage transformation can change the ordering of the eigenvalues and even make some of them negative. Stein suggested an *ad hoc* isotonizing algorithm that post-processes the transformed eigenvalues and thereby fixes these problems. We offer an alternative solution by minimizing the limiting expression of the unbiased estimator of risk under large-dimensional asymptotics, rather than the finite-sample expression. Compared to the isotonized version of Stein's estimator, our solution is theoretically more elegant and also delivers improved performance, as evidenced by Monte Carlo simulations.

KEY WORDS: Large-dimensional asymptotics, nonlinear shrinkage estimation, random matrix theory, rotation equivariance, Stein's loss.

JEL CLASSIFICATION NOS: C13.

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

1 Introduction

The estimation of a covariance matrix is one of the most fundamental problems in multivariate statistics. It has countless applications in econometrics, biostatistics, finance, signal processing, psychometrics, and many other fields. One recurrent problem is that the traditional estimator (that is, the sample covariance matrix) is ill-conditioned and performs poorly when the number of variables is not small compared to the sample size. Given the natural eagerness of applied researchers to look for patterns among as many variables as possible, and their practical ability to do so thanks to the ever-growing processing power of modern computers, theoreticians are under pressure to deliver estimation techniques that work well in large dimensions.

A famous proposal for improving over the sample covariance matrix in such cases is due to Stein (1975, 1986). He considers the class of *rotation-equivariant* estimators that keep the eigenvectors of the sample covariance matrix while shrinking its eigenvalues. This means that the small sample eigenvalues are pushed up and the large ones pulled down, thereby reducing (or *shrinking*) the overall spread of the set of eigenvalues. Stein's estimator is based on the scale-invariant loss function dating back to James and Stein (1961), and commonly referred to as *Stein's loss*.

Stein's shrinkage estimator broke new ground and fathered a large literature on rotation-equivariant shrinkage estimation of a covariance matrix. For example, see the works of Haff (1980), Lin and Perlman (1985), Dey and Srinivasan (1985), Daniels and Kaas (2001), Ledoit and Wolf (2004, 2012), Chen et al. (2009), Won et al. (2012), and the references therein.

Although to this day Stein's estimator has proven hard to surpass empirically, careful reading of the original article reveals a certain number of theoretical limitations.

1. The estimator proposed by Stein (1975, 1986) does not minimize the loss function, nor the expected loss (called the risk function), but an unbiased estimator of the risk. This is problematic because the primary objects of interest are the loss and the risk themselves. *A priori* there could exist many unbiased estimators of risk, minimizing them could lead to different estimators, and these estimators may or many not minimize the primary objects of interest: loss and/or risk.
2. The formula derived by Stein generates covariance matrix estimators that may not be positive semi-definite. To solve this problem, he recommends post-processing the estimator through an isotonizing algorithm. However, this is an *ad hoc* fix whose impact is not understood theoretically. In addition, the formula generates covariance matrix estimators that do not necessarily preserve the ordering of the eigenvalues of the sample covariance matrix. Once again, this problem forces the statistician to resort to the *ad hoc* isotonizing algorithm.
3. In order to derive his formula, Stein 'ignores' a term which involves the derivatives of the shrinkage function. No justification, apart from tractability, is given for this omission.
4. Finally, a more obvious limitation is that Stein's estimator requires normality, an assumption often violated by real data.

One important reason why Stein's estimator is highly regarded in spite of these four theoretical limitations is that several Monte Carlo simulations, such as the ones reported by Lin and Perlman (1985), have shown that it performs remarkably well in practice, as long as it is accompanied by the *ad hoc* isotonization algorithm.

Our paper develops a shrinkage estimator of the covariance matrix in the spirit of Stein (1975, 1986), with two significant improvements: first, it avoids the theoretical problems listed above; and second, it performs better in practice, as evidenced by extensive Monte-Carlo simulations. We respect Stein's framework by adopting Stein's loss as the metric by which estimators are evaluated, and by restricting ourselves to the class of rotation-equivariant estimators that have the same eigenvectors as the sample covariance matrix, like Stein does.

Our key innovation is to carry this framework from finite samples into the realm of *large-dimensional asymptotics*, where the number of variables and the sample size go to infinity together, with their ratio (called the *concentration*) converging to a finite, nonzero limit. Such an approach enables us to harness mathematical results from what is commonly known as *Random Matrix Theory* (RMT). It should be noted that Stein (1986) himself acknowledged the usefulness of RMT. He used it for illustration purposes only, which opens up the question of whether RMT could contribute more than that, and deliver a Stein-type estimator of the covariance matrix. Important new results in RMT enable us to answer positively.

We show that, under certain assumptions that are standard in the RMT literature, Stein's loss (properly normalized) converges almost surely to a nonrandom limit, which we characterize explicitly. We embed the eigenvalues of the covariance matrix estimator into a *shrinkage function*, and we introduce the notion of a *limiting* shrinkage function. The basic idea is that, even though the eigenvalues of the sample covariance matrix are random, the way they should be asymptotically transformed is nonrandom, and is governed by some limiting shrinkage function. We derive a necessary and sufficient condition for the limiting shrinkage function to minimize the large-dimensional asymptotic limit of Stein's loss. Finally, we construct a covariance matrix estimator that satisfies this condition, and thus is *asymptotically* optimal under Stein's loss in the class of rotation-equivariant estimators. Large-dimensional asymptotics enable us to:

1. show that Stein's loss function, the risk function, and Stein's unbiased estimator of risk are all asymptotically equivalent;
2. bypass the need for the isotonizing algorithm;
3. justify that the term involving the derivatives of the shrinkage function (which was ignored by Stein) vanishes indeed;
4. and relax the normality assumption.

These theoretical advantages translate into significantly improved practical performance over Stein's estimator, as we demonstrate through a comprehensive set of Monte Carlo simulations.

The paper is organized as follows. Section 2 briefly summarizes the finite-sample theory of Stein (1975, 1986). Section 3 details what adjustments are necessary to transplant Stein's

theory from finite samples to large-dimensional asymptotics. Section 4 develops our feasible estimator of a covariance matrix, which is asymptotically optimal. Section 5 studies finite-sample properties via Monte Carlo simulations. Section 6 contains concluding remarks. All mathematical proofs are collected in an appendix.

2 Shrinkage in Finite Samples under Stein's Loss

This section expounds the finite-sample theory of Stein (1975, 1986), with minor notational changes designed to enhance compatibility with the large-dimensional analysis conducted in subsequent sections. Such changes are highlighted where appropriate.

2.1 Finite-Sample Framework

Assumption 2.1 (Dimension)**.** *The number of variables p and the sample size n are both fixed and finite; p is smaller than n.*

Assumption 2.2 (Population Covariance Matrix)**.** *The population covariance matrix Σ_n is a nonrandom symmetric positive-definite matrix of dimension $p \times p$. Let $\boldsymbol{\tau}_n \coloneqq (\tau_{n,1}, \ldots, \tau_{n,p})'$ denote a system of eigenvalues of Σ_n. The empirical distribution function (e.d.f.) of the population eigenvalues is defined as: $\forall x \in \mathbb{R}$, $H_n(x) \coloneqq p^{-1} \sum_{i=1}^{p} \mathbb{1}_{[\tau_{n,i}, +\infty)}(x)$, where $\mathbb{1}$ denotes the indicator function of a set.*

Note that all relevant quantities are indexed by n because in subsequent sections we let the sample size n go to infinity (together with the dimension p).

Assumption 2.3 (Data Generating Process)**.** *X_n is a matrix of i.i.d. standard normal random variables of dimension $n \times p$. The matrix of observations is $Y_n \coloneqq X_n \times \sqrt{\Sigma_n}$, where $\sqrt{\Sigma_n}$ denotes the symmetric positive-definite square root of Σ_n. Neither $\sqrt{\Sigma_n}$ nor X_n are observed on their own: only Y_n is observed.*

The sample covariance matrix is defined as $S_n \coloneqq n^{-1} Y_n' Y_n = n^{-1} \sqrt{\Sigma_n} X_n' X_n \sqrt{\Sigma_n}$. It admits a spectral decomposition $S_n = U_n \Lambda_n U_n'$, where Λ_n is a diagonal matrix, and U_n is an orthogonal matrix: $U_n U_n' = U_n' U_n = \mathbb{I}_n$, where $\mathbb{I}_n$ (in slight abuse of notation) denotes the identity matrix of dimension $p \times p$. Let $\Lambda_n \coloneqq \mathsf{Diag}(\boldsymbol{\lambda}_n)$ where $\boldsymbol{\lambda}_n \coloneqq (\lambda_{n,1}, \ldots, \lambda_{n,p})'$. We can assume without loss of generality that the sample eigenvalues are sorted in increasing order: $\lambda_{n,1} \leq \lambda_{n,2} \leq \cdots \leq \lambda_{n,p}$. Correspondingly, the ith sample eigenvector is $u_{n,i}$, the ith column vector of U_n.

Assumption 2.4 (Estimators)**.** *We consider covariance matrix estimators of the type $\widetilde{S}_n \coloneqq U_n \widetilde{D}_n U_n'$, where $\widetilde{D}_n$ is a diagonal matrix: $\widetilde{D}_n \coloneqq \mathsf{Diag}\big(\lambda_{n,1} \widetilde{\psi}_n(\lambda_{n,1}) \ldots, \lambda_{n,p} \widetilde{\psi}_n(\lambda_{n,p})\big)$, and $\widetilde{\psi}_n$ is a (possibly random) real univariate function which can depend on S_n.*

This is the class of *rotation-equivariant* estimators introduced by Stein (1975, 1986): rotating the original variables results in the same rotation being applied to the covariance matrix

estimator. Such rotation equivariance is appropriate in the general case where the statistician has no *a priori* information about the orientation of the eigenvectors of the covariance matrix.

We call $\widetilde{\psi}_n$ the *shrinkage function* because, in all applications of interest, its effect is to shrink the set of sample eigenvalues by reducing its dispersion around the mean, pushing up the small ones and pulling down the large ones. Note that Stein (1986) does not work with the function $\widetilde{\psi}_n(\cdot)$ itself but with the vector $(\widetilde{\psi}_{n,1}, \ldots, \widetilde{\psi}_{n,p})' \coloneqq (\widetilde{\psi}_n(\lambda_{n,1}), \ldots, \widetilde{\psi}_n(\lambda_{n,p}))'$ instead. This is equivalent because the sample eigenvalues are distinct with probability one, and because the values taken by the shrinkage function $\widetilde{\psi}_n(\cdot)$ outside the set $\{\lambda_{n,1}, \ldots, \lambda_{n,p}\}$ do not make their way into the estimator $\widetilde{S}_n$. Of these two equivalent formulations, the functional one is easier to generalize into large-dimensional asymptotics than the vector one, for the same reason that authors in the Random Matrix Theory (RMT) literature have found it more tractable to work with the empirical distribution function (e.d.f.) of sample eigenvalues $F_n(x) \coloneqq p^{-1} \sum_{i=1}^{p} \mathbb{1}_{[\lambda_{n,i}, +\infty)}(x)$ than with the vector of sample eigenvalues.

Assumption 2.5 (Loss Function)**.** *Estimators are evaluated according to the following scale-invariant loss function used by Stein (1975, 1986) and commonly referred to as Stein's loss:*

$$\mathcal{L}_n(\Sigma_n, \widetilde{S}_n) \coloneqq \frac{1}{p} \mathsf{Tr}(\Sigma_n^{-1} \widetilde{S}_n) - \frac{1}{p} \log \det(\Sigma_n^{-1} \widetilde{S}_n) - 1 \ ,$$

and its corresponding risk function $\mathcal{R}_n(\Sigma_n, \widetilde{S}_n) \coloneqq \mathbb{E}[\mathcal{L}_n(\Sigma_n, \widetilde{S}_n)]$.

Note that Stein (1975, 1986) does not divide by p, but this normalization is necessary to prevent the loss function from going to infinity with the matrix dimension under large-dimensional asymptotics; it makes no difference in finite samples. By analogy with Stein's loss, we will refer to $\mathcal{R}_n(\Sigma_n, \widetilde{S}_n)$ as *Stein's risk*.

2.2 Stein's Loss in Finite Samples

Under Assumptions 2.1–2.5, Stein (1986) shows that the risk function verifies $\mathcal{R}_n(\Sigma_n, \widetilde{S}_n) = \mathbb{E}[\Theta_n(\Sigma_n, \widetilde{S}_n)]$, where

$$\begin{aligned} \Theta_n(\Sigma_n, \widetilde{S}_n) \coloneqq & \frac{n-p+1}{np} \sum_{j=1}^{p} \widetilde{\psi}_n(\lambda_{n,j}) - \frac{1}{p} \sum_{j=1}^{p} \log[\widetilde{\psi}_n(\lambda_{n,j})] + \log(n) \\ & + \frac{2}{np} \sum_{j=1}^{p} \sum_{i>j} \frac{\lambda_{n,j} \widetilde{\psi}_n(\lambda_{n,j}) - \lambda_{n,i} \widetilde{\psi}_n(\lambda_{n,i})}{\lambda_{n,j} - \lambda_{n,i}} \\ & + \frac{2}{np} \sum_{j=1}^{p} \lambda_{n,j} \widetilde{\psi}_n'(\lambda_{n,j}) - \frac{1}{p} \sum_{j=1}^{p} \mathbb{E}[\log(\chi^2_{n-j+1})] - 1 \ , \end{aligned} \tag{2.1}$$

with

$$\widetilde{\psi}_n'(x) \coloneqq \frac{\partial \widetilde{\psi}_n(x)}{\partial x} \ .$$

Therefore, the random quantity $\Theta_n(\Sigma_n, \widetilde{S}_n)$ can be interpreted as an *unbiased estimator of the risk (function)*.

Ignoring the term $(2/np)\sum_{j=1}^{p}\lambda_{n,j}\widetilde{\psi}'_n(\lambda_j)$, the unbiased estimator of risk is minimized when the shrinkage function $\widetilde{\psi}_n$ satisfies $\forall i=1,\ldots,p,\ \ \widetilde{\psi}_n(\lambda_{n,i})=\psi^*_n(\lambda_{n,i})$, where

$$\forall i=1,\ldots,p \qquad \psi^*_n(\lambda_{n,i}) \coloneqq \frac{1}{1-\dfrac{p-1}{n}-2\,\dfrac{p}{n}\,\lambda_{n,i}\times\dfrac{1}{p}\displaystyle\sum_{j\neq i}\dfrac{1}{\lambda_{n,j}-\lambda_{n,i}}}\ . \tag{2.2}$$

While this approach broke new ground and had a major impact on subsequent developments in multivariate statistics, a drawback of working in finite samples is that expression (2.2) diverges when some $\lambda_{n,j}$ gets infinitesimally close to another $\lambda_{n,i}$. In such cases, Stein's original estimator can exhibit violation of eigenvalues ordering, or even negative eigenvalues. It necessitates post-processing through an *ad hoc* isotonizing algorithm whose effect is hard to quantify theoretically. This is one of the motivations for going to large-dimensional asymptotics.

The appendix of Lin and Perlman (1985) gives a detailed description of the isotonizing algorithm. If we call the isotonized shrinkage function ψ^{ST}_n, Stein's *isotonized* estimator is

$$S^{ST}_n \coloneqq U_n D^{ST}_n U'_n \quad \text{where} \quad D^{ST}_n \coloneqq \mathsf{Diag}(\lambda_{n,1}\psi^{ST}_n(\lambda_{n,1}),\ldots,\lambda_{n,p}\psi^{ST}_n(\lambda_{n,p}))\ . \tag{2.3}$$

3 Shrinkage in Large Dimensions under Stein's Loss

This section largely mirrors the previous one, and contains adjustments designed to convert from finite samples to large-dimensional asymptotics.

3.1 Large-Dimensional Asymptotic Framework

Assumption 3.1 (Dimension)**.** *Let n denote the sample size and $p\coloneqq p(n)$ the number of variables. It is assumed that the ratio p/n converges, as $n\to\infty$, to a limit $c\in(0,1)$ called the* concentration*. Furthermore, there exists a compact interval included in $(0,1)$ that contains p/n for all n large enough.*

Assumption 3.2 (Population Covariance Matrix)**.** *The population covariance matrix Σ_n is a nonrandom symmetric positive-definite matrix of dimension $p\times p$. Let $\boldsymbol{\tau}_n\coloneqq(\tau_{n,1},\ldots,\tau_{n,p})'$ denote a system of eigenvalues of Σ_n, and H_n the e.d.f. of population eigenvalues. It is assumed that H_n converges weakly to a limit law H, called the limiting spectral distribution (function). $\mathsf{Supp}(H)$, 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 $[\underline{h},\overline{h}]\subset(0,\infty)$ that contains $\mathsf{Supp}(H_n)$ for all n large enough.*

The existence of a limiting concentration (ratio) and a limiting population spectral distribution are standard assumptions in the literature on large-dimensional asymptotics; see Bai and Silverstein (2010) for a comprehensive review.

Assumption 3.3 (Data Generating Process)**.** *X_n is an $n\times p$ matrix of i.i.d. random variables with mean zero, variance one, and finite 12th moment. The matrix of observations is $Y_n\coloneqq$*

$X_n \times \sqrt{\Sigma_n}$, *where* $\sqrt{\Sigma_n}$ *denotes the symmetric positive-definite square root of* Σ_n. *Neither* $\sqrt{\Sigma_n}$ *nor* X_n *are observed on their own: only* Y_n *is observed.*

Note that we no longer require normality.

The literature on sample covariance matrix eigenvalues under large-dimensional asymptotics is based on a foundational result by Marčenko and Pastur (1967). It has been strengthened and broadened by subsequent authors including Silverstein (1995), Silverstein and Bai (1995), and Silverstein and Choi (1995), among others. These articles imply that, under Assumptions 3.1–3.3, there exists a continuously differentiable limiting sample spectral distribution F such that

$$\forall x \in \mathbb{R} \qquad F_n(x) - F(x) \xrightarrow{\text{a.s.}} 0 \ . \tag{3.1}$$

In addition, the existing literature has unearthed important information about the limiting spectral distribution F, including an equation that relates F to H and c. The version of this equation given by Silverstein (1995) is that $m \coloneqq m_F(z)$ is the unique solution in the set

$$\left\{ m \in \mathbb{C} : -\frac{1-c}{z} + cm \in \mathbb{C}^+ \right\} \tag{3.2}$$

to the equation

$$\forall z \in \mathbb{C}^+ \qquad m_F(z) = \int \frac{1}{\tau\big[1 - c - c\,z\,m_F(z)\big] - z}\, dH(\tau) \ , \tag{3.3}$$

where $\mathbb{C}^+$ is the half-plane of complex numbers with strictly positive imaginary part and, for any increasing function G on the real line, m_G denotes the Stieltjes transform of G:

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

The Stieltjes transform admits a well-known inversion formula:

$$G(b) - G(a) = \lim_{\eta \to 0^+} \frac{1}{\pi} \int_a^b \mathsf{Im}\big[m_G(\xi + i\eta)\big] d\xi \ , \tag{3.4}$$

if G is continuous at a and b. While the Stieltjes transform of F, m_F, is a function whose domain is the upper half of the complex plane, it admits an extension to the real line, since Silverstein and Choi (1995) show that: $\forall \lambda \in \mathbb{R}$, $\lim_{z \in \mathbb{C}^+ \to \lambda} m_F(z) =: \breve{m}_F(\lambda)$ exists and is continuous.

Another useful result concerns the support of the distribution of sample eigenvalues. Theorem 1.1 of Bai and Silverstein (1998) and Assumptions 3.1–3.3 imply that the support of F, $\mathsf{Supp}(F)$, is the union of a finite number $\kappa \geq 1$ of compact intervals: $\mathsf{Supp}(F) = \bigcup_{k=1}^{\kappa} [a_k, b_k]$, where $0 < a_1 < b_1 < \cdots < a_\kappa < b_\kappa < \infty$.

Assumption 3.4 (Estimators)**.** *We consider covariance matrix estimators of the type* $\widetilde{S}_n \coloneqq U_n \widetilde{D}_n U_n'$ *where* $\widetilde{D}_n$ *is a diagonal matrix:* $\widetilde{D}_n \coloneqq \mathsf{Diag}\big(\lambda_{n,i}\widetilde{\psi}_n(\lambda_{n,i}) \ldots, \lambda_{n,i}\widetilde{\psi}_n(\lambda_{n,i})\big)$, *and* $\widetilde{\psi}_n$ *is a (possibly random) real univariate function which can depend on* S_n. *We assume that there exists*

a nonrandom real univariate function $\widetilde{\psi}$ defined on $\mathsf{Supp}(F)$ and continuously differentiable such that $\widetilde{\psi}_n(x) \xrightarrow{\text{a.s}} \widetilde{\psi}(x)$ for all $x \in \mathsf{Supp}(F)$. Furthermore, this convergence is uniform over $x \in \bigcup_{k=1}^{\kappa}[a_k + \eta, b_k - \eta]$, for any small $\eta > 0$. Finally, for any small $\eta > 0$, there exists a finite nonrandom constant $\widetilde{K}$ such that almost surely, over the set $x \in \bigcup_{k=1}^{\kappa}[a_k - \eta, b_k + \eta]$, $|\widetilde{\psi}_n(x)|$ is uniformly bounded by $\widetilde{K}$, for all n large enough.

Shrinkage functions need to be as well behaved asymptotically as spectral distribution functions, except possibly on a finite number of arbitrarily small regions near the boundary of the support. The large-dimensional asymptotic properties of a generic rotation-equivariant estimator $\widetilde{S}_n$ are fully characterized by its limiting shrinkage function $\widetilde{\psi}$.

Assumption 3.5 (Loss Function)**.** *Estimators are evaluated according to the limit, as n and p go to infinity together, of the following loss function:*

$$\mathcal{L}_n(\Sigma_n, \widetilde{S}_n) := \frac{1}{p}\mathsf{Tr}(\Sigma_n^{-1}\widetilde{S}_n) - \frac{1}{p}\log\,\det(\Sigma_n^{-1}\widetilde{S}_n) - 1\ ,$$

and of its corresponding risk function $\mathcal{R}_n(\Sigma_n, \widetilde{S}_n) := \mathbb{E}[\mathcal{L}_n(\Sigma_n, \widetilde{S}_n)]$.

The key difference is that, instead of minimizing the unbiased estimator of risk $\Theta_n(\Sigma_n, \widetilde{S}_n)$ from equation (2.1), as Stein (1986) does, we minimize $\lim_{n,p\to\infty}\Theta_n(\Sigma_n, \widetilde{S}_n)$. The almost sure existence of this limit is established below.

3.2 Stein's Loss under Large-Dimensional Asymptotics

Theorem 3.1. *Under Assumptions 3.1–3.5,*

$$\begin{aligned}\mathcal{L}_n(\Sigma_n, \widetilde{S}_n) \xrightarrow{\text{a.s.}} \sum_{k=1}^{\kappa}\int_{a_k}^{b_k}\Big\{\big(1 - c - 2\,c\,x\,\mathsf{Re}[\breve{m}_F(x)]\big)\widetilde{\psi}(x) - \log[\widetilde{\psi}(x)]\Big\}dF(x)\\ + \frac{1-c}{c}\log(1-c)\ .\end{aligned} \tag{3.5}$$

The proof is in Appendix A. The connection with Stein's finite sample-analysis is further elucidated by an equivalent result for the unbiased estimator of risk.

Proposition 3.1. *Under Assumptions 3.1–3.5,*

$$\begin{aligned}\Theta_n(\Sigma_n, \widetilde{S}_n) \xrightarrow{\text{a.s.}} \sum_{k=1}^{\kappa}\int_{a_k}^{b_k}\Big\{\big(1 - c - 2\,c\,x\,\mathsf{Re}[\breve{m}_F(x)]\big)\widetilde{\psi}(x) - \log[\widetilde{\psi}(x)]\Big\}dF(x)\\ + \frac{1-c}{c}\log(1-c)\ .\end{aligned} \tag{3.6}$$

The proof is in Appendix B. Proposition 3.1 shows that, under large-dimensional asymptotics, minimizing the unbiased estimator of risk is actually equivalent to minimizing the loss, with probability one. It also shows that ignoring the term $(2/np)\sum_{j=1}^{p}\lambda_{n,j}\widetilde{\psi}'_n(\lambda_j)$ in the unbiased estimator of risk, which was an *ad hoc* approximation in finite samples, is justified under large-dimensional asymptotics, since this term vanishes in the limit.

Theorem 3.1 enables us to characterize the set of asymptotically optimal estimators under Stein's loss in large dimensions.

Corollary 3.1. *Under Assumptions 3.1–3.5, a covariance matrix estimator $\widetilde{S}_n$ minimizes the almost sure limit of Stein's loss if and only if its limiting shrinkage function $\widetilde{\psi}$ verifies $\forall x \in \mathsf{Supp}(F)$ $\widetilde{\psi}(x) = \psi^*(x)$, where*

$$\forall x \in \mathsf{Supp}(F) \qquad \psi^*(x) \coloneqq \frac{1}{1 - c - 2\,c\,x\,\mathsf{Re}[\breve{m}_F(x)]} \ . \tag{3.7}$$

The proof follows immediately from Theorem 3.1 by differentiating the right-hand side of equation (3.5) with respect to $\widetilde{\psi}(x)$.

Remark 3.1 (Stein's loss versus Frobenius-norm loss)**.** Interestingly, an equivalent formula to equation (3.7) is attained by Ledoit and Péché (2011, Theorem 5), even though it is motivated by a different loss function, namely $\|\Sigma_n^{-1} - \widehat{\Sigma}_n^{-1}\|_F$, where $\|A\|_F \coloneqq \mathsf{Tr}(AA')$ denotes the Frobenius norm of a matrix. The formula of Ledoit and Péché (2011) is the reciprocal of equation (3.7), as the object of interest is the *precision matrix* (that is, the inverse of the covariance matrix) instead of the covariance matrix itself. The loss function $\|\Sigma_n - \widehat{\Sigma}_n\|_F$ leads to a different asymptotic formula, as shown by Ledoit and Péché (2011, Theorem 4).

The fact that the denominator on the right-hand side of equation (3.7) is nonzero and that the optimal limiting shrinkage function ψ^* is strictly positive and bounded over the support of F is established by the following proposition.

Proposition 3.2. *Under Assumptions 3.1–3.3*

$$\forall x \in \mathsf{Supp}(F) \qquad 1 - c - 2\,c\,x\,\mathsf{Re}[\breve{m}_F(x)] \geq \frac{a_1}{\overline{h}} \ .$$

The proof is in Appendix C.

4 Optimal Covariance Matrix Estimation under Stein's Loss

The Stieltjes transform of the limiting distribution of sample eigenvalues F contained in formula (3.7) is unobservable. Therefore, formula (3.7) can be interpreted as an infeasible *oracle*. Ledoit and Wolf (2013), extending the methodology of Ledoit and Wolf (2012), develop an approach that can be used to estimate this Stieltjes transform consistently.

4.1 The QuEST Function

Ledoit and Wolf (2013) 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 (3.1)–(3.4). For any positive integers n and p, the QuEST function, denoted by $Q_{n,p}$, is defined as

$$Q_{n,p} : [0,\infty)^p \longrightarrow [0,\infty)^p \tag{4.1}$$

$$\mathbf{t} \coloneqq (t_1, \ldots, t_p)' \longmapsto Q_{n,p}(\mathbf{t}) \coloneqq \left(q_{n,p}^1(\mathbf{t}), \ldots, q_{n,p}^p(\mathbf{t})\right)' \ , \tag{4.2}$$

where

$$\forall i = 1, \ldots, p \qquad q_{n,p}^{i}(\mathbf{t}) \coloneqq p \int_{(i-1)/p}^{i/p} \left(F_{n,p}^{\mathbf{t}}\right)^{-1}(u)\, du \ , \tag{4.3}$$

$$\forall u \in [0,1] \qquad \left(F_{n,p}^{\mathbf{t}}\right)^{-1}(u) \coloneqq \sup\{x \in \mathbb{R} : F_{n,p}^{\mathbf{t}}(x) \leq u\} \ , \tag{4.4}$$

$$\forall x \in \mathbb{R} \qquad F_{n,p}^{\mathbf{t}}(x) \coloneqq \lim_{\eta \to 0^+} \frac{1}{\pi} \int_{-\infty}^{x} \mathsf{Im}\left[m_{n,p}^{\mathbf{t}}(\xi + i\eta)\right] d\xi \ , \tag{4.5}$$

and $\forall z \in \mathbb{C}^+ \quad m \coloneqq m_{n,p}^{\mathbf{t}}(z)$ is the unique solution in the set

$$\left\{ m \in \mathbb{C} : -\frac{n-p}{nz} + \frac{p}{n}\, m \in \mathbb{C}^+ \right\} \tag{4.6}$$

to the equation

$$m = \frac{1}{p} \sum_{i=1}^{p} \frac{1}{t_i \left(1 - \dfrac{p}{n} - \dfrac{p}{n}\, z\, m\right) - z} \ . \tag{4.7}$$

It can be seen that equation (4.5) quantizes equation (3.4), that equation (4.6) quantizes equation (3.2), and that equation (4.7) quantizes equation (3.3). Thus, $F_{n,p}^{\mathbf{t}}$ 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 (4.4), $\left(F_{n,p}^{\mathbf{t}}\right)^{-1}$ represents the inverse spectral distribution function, also known as the *quantile* function. By equation (4.3), $q_{n,p}^{i}(\mathbf{t})$ can be interpreted as a 'smoothed' version of the $(i - 0.5)/p$ quantile of $F_{n,p}^{\mathbf{t}}$.

4.2 Consistent Estimator of Population Eigenvalues

Ledoit and Wolf (2013) estimate the eigenvalues of the population covariance matrix by numerically inverting the QuEST function.

Theorem 4.1. *Suppose that Assumptions 3.1–3.3 are satisfied. Define*

$$\widehat{\boldsymbol{\tau}}_n \coloneqq \underset{\mathbf{t} \in (0,\infty)^p}{\operatorname{argmin}} \frac{1}{p} \sum_{i=1}^{p} \left[q_{n,p}^{i}(\mathbf{t}) - \lambda_{n,i}\right]^2 , \tag{4.8}$$

where $\boldsymbol{\lambda}_n \coloneqq (\lambda_{n,1}, \ldots, \lambda_{n,p})'$ *are the sample covariance matrix eigenvalues, and* $Q_{n,p}(\mathbf{t}) \coloneqq \left(q_{n,p}^{1}(\mathbf{t}), \ldots, q_{n,p}^{p}(\mathbf{t})\right)'$ *is the nonrandom QuEST function defined in equations* (4.1)–(4.7)*; both* $\widehat{\boldsymbol{\tau}}_n$ *and* $\boldsymbol{\lambda}_n$ *are assumed sorted in nondecreasing order. Let* $\widehat{\tau}_{n,i}$ *denote the* i*th entry of* $\widehat{\boldsymbol{\tau}}_n$ $(i = 1, \ldots, p)$*, and let* $\boldsymbol{\tau}_n \coloneqq (\tau_{n,1}, \ldots, \tau_{n,p})'$ *denote the population covariance matrix eigenvalues sorted in nondecreasing order. Then*

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

The proof is given by Ledoit and Wolf (2013, Theorem 2.2). The solution to equation (4.8) can be found by standard nonlinear optimization software such as SNOPT$^{\text{TM}}$; see Gill et al. (2002).

4.3 Asymptotically Optimal Estimator of the Covariance Matrix

Recall that, for any $\mathbf{t} := (t_1, \ldots, t_p)' \in (0, +\infty)^p$, equations (4.6)–(4.7) define $m_{n,p}^{\mathbf{t}}$ as the Stieltjes transform of $F_{n,p}^{\mathbf{t}}$, the limiting distribution function of sample eigenvalues corresponding to the population spectral distribution function $p^{-1} \sum_{i=1}^{p} \mathbb{1}_{[t_i, +\infty)}$. The domain of $m_{n,p}^{\mathbf{t}}$ 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 $\forall \lambda \in \mathbb{R} \quad \lim_{z \in \mathbb{C}^+ \to \lambda} m_{n,p}^{\mathbf{t}}(z) =: \breve{m}_{n,p}^{\mathbf{t}}(\lambda)$ exists. An asymptotically optimal estimator of the covariance matrix can be constructed simply by plugging into equation (3.7) the estimator of the population eigenvalues obtained in equation (4.8).

Theorem 4.2. *Suppose that Assumptions 3.1–3.5 are satisfied. The covariance matrix estimator defined by*

$$\begin{aligned} &\widehat{S}_n := U_n \widehat{D}_n U_n' \quad \textit{where} \quad \widehat{D}_n := \mathsf{Diag}(\lambda_{n,1} \widehat{\psi}_n(\lambda_{n,1}), \ldots, \lambda_{n,p} \widehat{\psi}_n(\lambda_{n,p})) \\ &\textit{and} \quad \forall i = 1, \ldots, p \quad \widehat{\psi}_n(\lambda_{n,i}) := \frac{1}{1 - \dfrac{p}{n} - 2 \dfrac{p}{n} \lambda_{n,i} \, \mathsf{Re}\big[\breve{m}_{n,p}^{\widehat{\boldsymbol{\tau}}_n}(\lambda_{n,i})\big]} \end{aligned} \tag{4.9}$$

minimizes in the class of rotation-equivariant estimators described in Assumption 2.4 the almost sure limit of Stein's loss as n and p go to infinity together.

The proof is in Appendix D. The structure of formula (4.9) is very similar to the structure of the corresponding formula in Stein (1975, 1986). The key difference lies in the term $\mathsf{Re}\big[\breve{m}_{n,p}^{\widehat{\boldsymbol{\tau}}_n}(\lambda_{n,i})\big]$. In its place, equation (2.2) has

$$\frac{1}{p} \sum_{i \neq j} \frac{1}{\lambda_{n,j} - \lambda_{n,i}} \, .$$

The connection is that $\mathsf{Re}\big[\breve{m}_{n,p}^{\widehat{\boldsymbol{\tau}}_n}(\lambda_{n,i})\big]$ can be expressed as a (Cauchy) Principal Value.

$$\begin{aligned} \mathsf{Re}\big[\breve{m}_{n,p}^{\widehat{\boldsymbol{\tau}}_n}(\lambda_{n,i})\big] &= \mathrm{PV}\left[\int \frac{1}{\lambda - \lambda_{n,i}} dF_{n,p}^{\widehat{\boldsymbol{\tau}}_n}(\lambda)\right] \\ &:= \lim_{\varepsilon \searrow 0} \left\{ \int_{-\infty}^{\lambda_{n,i} - \varepsilon} \frac{1}{\lambda - \lambda_{n,i}} dF_{n,p}^{\widehat{\boldsymbol{\tau}}_n}(\lambda) + \int_{\lambda_{n,i} + \varepsilon}^{\infty} \frac{1}{\lambda - \lambda_{n,i}} dF_{n,p}^{\widehat{\boldsymbol{\tau}}_n}(\lambda) \right\} . \end{aligned}$$

See, e.g., Henrici (1988, pp. 259–262) for a reference on Principal Values. While Stein's original estimator picks the $\lambda_{n,j}$'s from the empirical distribution function of the sample eigenvalues, F_n, which is a step function, our estimator picks the λ's from the smooth distribution $F_{n,p}^{\widehat{\boldsymbol{\tau}}_n}$ instead. This enables us to avoid the problems that beset Stein's original estimator when some $\lambda_{n,j}$ happens to be 'too close' to another $\lambda_{n,i}$, such as violation of eigenvalue ordering or negative eigenvalues.

Remark 4.1 (Stein's loss versus Frobenius-norm loss; Remark 3.1 continued)**.** The loss function $\|\Sigma_n - \widehat{\Sigma}_n\|_F$ leads to a different asymptotically optimal estimator of the covariance matrix, as discussed by Ledoit and Wolf (2012, 2013). Which loss function, Stein's loss or Frobenius-norm loss, is more appropriate may generally depend on the objective for which the estimator

of the covariance matrix is intended (construction of a test statistic, linear discriminant analysis, efficient generalized method of moments, Markowitz-type portfolio optimization, etc.). It is not the goal of this paper to enter such a debate. On the other hand, we would like to point out that the *same* methodology — using the approach of Ledoit and Wolf (2013) to consistently estimate the oracle formulæ of Ledoit and Péché (2011) — ultimately delivers the asymptotically optimal solution for *both* loss functions.

5 Monte Carlo Simulations

For compactness of notation, in this section, "Stein's estimator" stands for "Stein's isotonized estimator" always.

The isotonized shrinkage estimator of Stein (1986) is widely acknowledged to have very good performance in Monte Carlo simulations, which compensates for theoretical limitations such as the recourse to an *ad hoc* isotonizing algorithm, minimizing an unbiased estimator of risk instead of the risk itself, and neglecting the derivatives term in equation (2.1). The article by Lin and Perlman (1985) is a prime example of the success of Stein's estimator in Monte Carlo simulations.

We report a set of Monte Carlo simulations comparing the nonlinear shrinkage estimator developed in Theorem 4.2 with Stein's estimator. There exist a host of alternative rotation-equivariant shrinkage estimators of a covariance matrix; see the literature review in the introduction. Including all of them in the Monte Carlo simulations is certainly beyond the scope of the paper.

The chosen metric is the Percentage Relative Improvement in Average Loss (PRIAL) relative to Stein's estimator. For a generic estimator $\widehat{\Sigma}_n$, define

$$\text{PRIAL}(S_n^{ST}, \widehat{\Sigma}_n) \coloneqq \left[1 - \frac{\mathcal{R}_n(\Sigma_n, \widehat{\Sigma}_n)}{\mathcal{R}_n(\Sigma_n, S_n^{ST})}\right] \times 100\% \; .$$

Thus $\text{PRIAL}(S_n^{ST}, S_n^{ST}) = 0\%$ and $\text{PRIAL}(S_n^{ST}, \Sigma_n) = 100\%$ by construction. We report $\text{PRIAL}(S_n^{ST}, \widehat{S}_n)$, where the empirical risks of S_n^{ST} and $\widehat{S}_n$ are computed as averages across 1,000 Monte Carlo simulations.

In all designs, the ith population eigenvalue is equal to $\tau_{n,i} \equiv H^{-1}((i-0.5)/p)$ $(i = 1, \ldots, p)$, where H is the limiting population spectral distribution. Unless stated otherwise, the distribution of the random variates comprising the $n \times p$ data matrix X_n is Gaussian.

Our numerical experiments are built around a 'baseline' scenario, and we vary different design elements in turn. In the baseline case, $p = 100$, $n = 200$, and H is given by the distribution of $1 + W$, where $W \sim \text{Beta}(2, 5)$. This distribution is right-skewed, meaning that there are a lot of small eigenvalues and few big ones, which is representative of many practically relevant situations; see Figure 4 below. In this case, the PRIAL of our new nonlinear shrinkage estimator relative to Stein's is 42%.

Convergence

First, we vary the matrix dimension p from $p = 30$ to $p = 200$ while keeping the concentration ratio p/n fixed at the value $1/2$. The results are displayed in Figure 1.

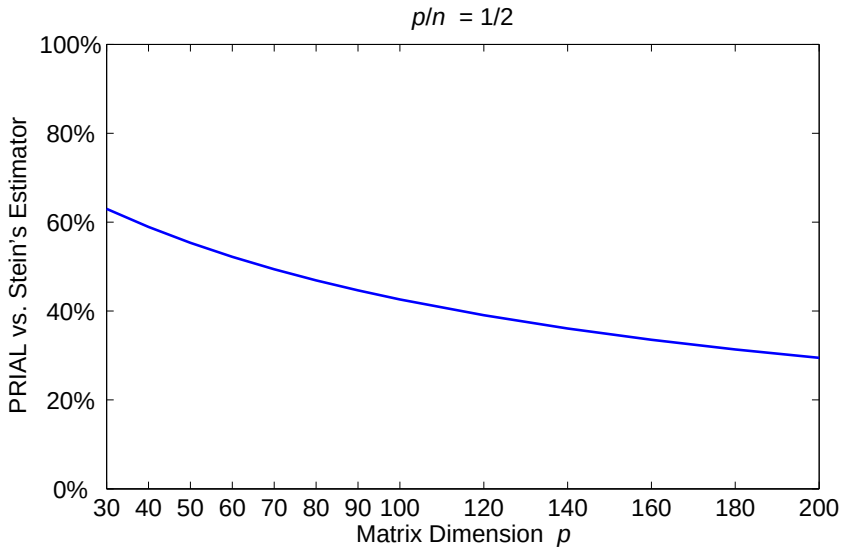

Figure 1: Evolution of the PRIAL of the new nonlinear shrinkage estimator relative to Stein's isotonized estimator as matrix dimension and sample size go to infinity together.

One can see that the improvement is strong across the board, and stronger in small-to-medium dimensions.

Concentration

Second, we vary the concentration (ratio) from $p/n = 0.05$ to $p/n = 0.94$ while keeping the product $p \times n$ constant at the value $20,000$. The results are displayed in Figure 2.

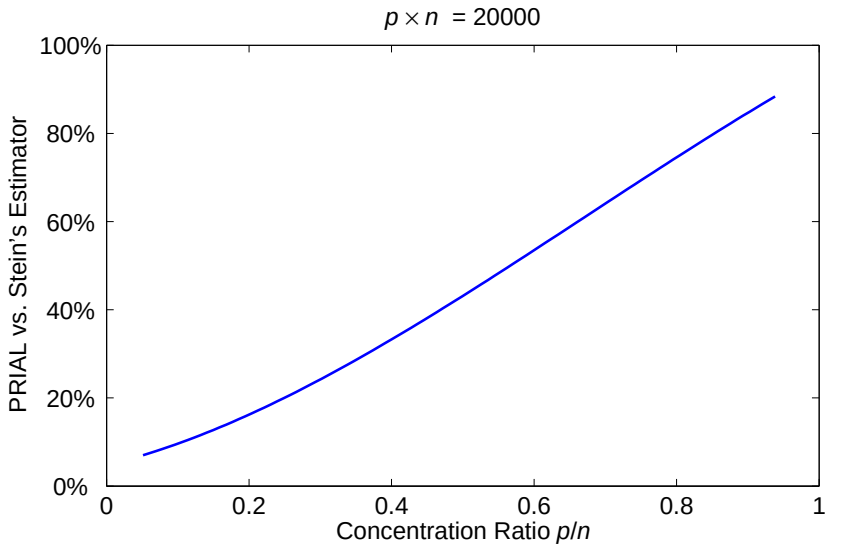

Figure 2: PRIAL of the new nonlinear shrinkage estimator relative to Stein's isotonized estimator as a function of the concentration ratio p/n.

One can that the improvement is good across the board, and stronger when the matrix dimension is close to the sample size.

Condition Number

Third, we vary the condition number of the population covariance matrix. We do this by taking H to be the distribution of $a + (2 - a)W$, where $W \sim \text{Beta}(2, 5)$. Across all values of $a \in [0.01, 2]$, the upper bound of the support of H remains constant at the value 2, while the lower bound of the support is equal to a. Consequently, the condition number decreases in a from 32 to 1. The results are displayed in Figure 3.

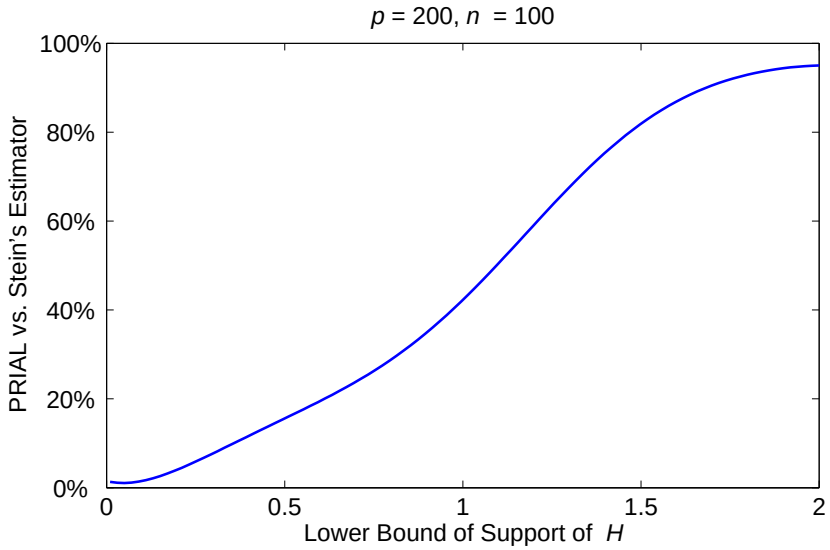

Figure 3: PRIAL of the new nonlinear shrinkage estimator relative to Stein's isotonized estimator across various condition numbers.

One can see that the improvement is positive across the board, and increases as the population covariance matrix becomes better conditioned.

Shape

Fourth, we vary the shape of the distribution of the population eigenvalues. We take H to be the distribution of $1 + W$, where $W \sim \text{Beta}(\alpha, \beta)$ for various pairs of parameters (α, β). The corresponding densities are displayed in Figure 4.

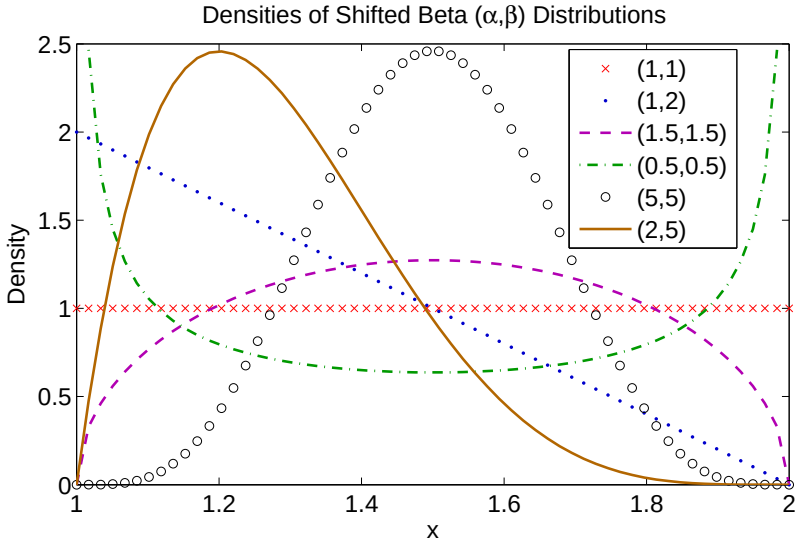

Figure 4: Densities of various shifted Beta distributions. Note that the density of Beta(β, α) is just the mirror image (around the mid point of the support) of the density of Beta(α, β).

The results are presented in Table 1.

Parameters	PRIAL
$(1, 1)$	21%
$(1, 2)$	27%
$(2, 1)$	31%
$(1.5, 1.5)$	26%
$(0.5, 0.5)$	15%
$(5, 5)$	52%
$(2, 5)$	42%
$(5, 2)$	55%
Average	34%

Table 1: PRIAL of the nonlinear shrinkage estimator relative to Stein's isotonized estimator for various shapes of the population spectral distribution.

There is no obvious pattern; the improvement is good across all distribution shapes, and the baseline case $(\alpha, \beta) = (2, 5)$ is neither the best nor the worst.

Non-normality

Finally, we vary the distribution of the variates X_n. Beyond the normal distribution with kurtosis 0, we also consider the coin-toss Bernoulli distribution, which is platykurtic with kurtosis -2, and the Laplace distribution, which is leptokurtic with kurtosis 3. The results are presented in Table 2.

Distribution	PRIAL
Normal	42%
Bernoulli	42%
Laplace	44%

Table 2: PRIAL of the nonlinear shrinkage estimator relative to Stein's isotonized estimator for various distributions of the variates.

One can see that the results obtained above carry over to the non-normal case.

Overall, the conclusion from these numerical experiments is that, even though Stein's estimator is known for performing very well in Monte Carlo simulations, our new nonlinear shrinkage estimator improves substantially over it across a wide variety of situations. The improvement is strongest when the sample size is not very large and the population eigenvalues are not very dispersed.

6 Concluding Remarks

Estimating a covariance matrix is one of the most fundamental problems in statistics, with a host of important applications. But in a large-dimensional setting, when the number of variables is not small compared to the sample size, the traditional estimator (that is, the sample covariance matrix) is ill-conditioned and performs poorly.

This paper has revisited the pioneering work of Stein (1975, 1986) to construct an improved estimator of a covariance matrix, based on a scale-invariant loss function commonly known as Stein's loss. The estimator originally proposed by Stein suffers from two pitfalls: violation of eigenvalue ordering and the possibility of negative eigenvalues (that is, a negative definite estimator of a covariance matrix). As a dual remedy, Stein proposed an *ad hoc* isotonizing algorithm to be applied to the eigenvalues of his original estimator.

Stein derived has original estimator by minimizing an unbiased estimator of risk in finite samples, considering a certain class of rotation-equivariant estimators (and assuming multivariate normality). In contrast, we have opted for large-dimensional asymptotic analysis, considering the same class of rotation-equivariant estimators. We show that the unbiased estimator of risk for such an estimator, under mild regularity conditions (where even the assumption of multivariate normality can be dropped), almost surely converges to a nonrandom limit; and that this limit is actually equal to the almost sure limit of the value of the loss. Our alternative estimator is then based on minimizing this limiting expression of the loss.

Not surprisingly, the initial solution depends on unknown population quantities, resulting in an oracle estimator. However, using recent nonlinear shrinkage methodology developed by Ledoit and Wolf (2012, 2013) with tools from the Random Matrix Theory literature, we are

able to derive a *bona fide* estimator that is also asymptotically optimal (in the sense of minimizing the limiting expression of the loss). This enables us to avoid the theoretical difficulties that beset Stein's estimator and also to improve finite-sample performance, as evidenced by extensive simulation studies.

References

Abramowitz, M. and Stegun, I. A. (1965). *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*, volume 55. Dover publications.

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. (2004). CLT for linear spectral statistics of large-dimensional sample covariance matrices. *The Annals of Probability*, 32(1A):553–605.

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

Chen, Y., Wiesel, A., and Hero, A. O. (2009). Shrinkage estimation of high dimensional covariance matrices. IEEE International Conference on Acoustics, Speech, and Signal Processing, Taiwan.

Daniels, M. J. and Kaas, R. E. (2001). Shrinkage estimators for covariance matrices. *Biometrics*, 57:1173–1184.

Dey, D. K. and Srinivasan, C. (1985). Estimation of a covariance matrix under Stein's loss. *Annals of Statistics*, 13(4):1581–1591.

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.

Haff, L. R. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix. *Annals of Statistics*, 8:586–597.

Henrici, P. (1988). *Applied and Computational Complex Analysis*, volume 1. Wiley, New York.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* **1**, pages 361–380.

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.

Ledoit, O. and Wolf, M. (2013). Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions. Working Paper ECON 105, Department of Economics, University of Zurich.

Lin, S. and Perlman, M. D. (1985). A Monte Carlo comparison of four estimators of a covariance matrix. *Multivariate Analysis*, 6:411–429.

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

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.

Won, J.-H., Lim, J., Kim, S.-J., and Rajaratnam, B. (2012). Condition-number regularized covariance estimation. *Journal of the Royal Statistical Society B*, 75(3).

Appendix

For notational simplicity, the proofs below assume that the support of F is a single compact interval $[a,b] \subset (0,+\infty)$. But they generalize easily to the case where $\mathsf{Supp}(F)$ is the union of a finite number κ of such intervals, as maintained in Assumptions 3.2 and 3.4.

When there is no ambiguity, the first subscript, n, can be dropped from the notation of the eigenvalues and eigenvectors.

A Proof of Theorem 3.1

Lemma A.1. *Define* $\forall x \in \mathbb{R} \quad \Phi_n(x) := p^{-1} \sum_{i=1}^{p} u_i' \Sigma_n^{-1} u_i \times \mathbb{1}_{[\lambda_i, +\infty)}(x)$. *Under Assumptions 3.1–3.3, there exists a nonrandom function* Φ *defined on* $\mathbb{R}$ *such that* $\Phi_n(x)$ *converges almost surely to* $\Phi(x)$*, for all* $x \in \mathbb{R}$*. Furthermore,* Φ *is continuously differentiable on* $\mathbb{R}$ *and can be expressed as*

$$\forall x \in \mathbb{R} \qquad \Phi(x) = \begin{cases} 0 & \text{if } x < a, \\ \int_a^x \varphi(\lambda) dF(\lambda) & \text{if } x \geq a, \end{cases}$$

where $\forall \lambda \in [a, +\infty) \;\; \varphi(\lambda) := \big\{1 - c - 2\,c\,\lambda\,\mathsf{Re}[\breve{m}_F(\lambda)]\big\}/\lambda$.

Proof of Lemma A.1. The proof of Lemma A.1 follows directly from Ledoit and Péché (2011, Theorem 5) and the corresponding proof, bearing in mind that we are in the case $c < 1$ because of Assumption 3.1. ■

Lemma A.2. *Under Assumptions (3.1)–(3.4),*

$$\frac{1}{p} \mathsf{Tr}\big(\Sigma_n^{-1} \widetilde{S}_n\big) \xrightarrow{\text{a.s.}} \int_a^b x\, \widetilde{\psi}(x)\, d\Phi(x) \;.$$

Proof of Lemma A.2. Restrict attention to the set Ω_1 of probability one on which $\Phi_n(x)$ converges to $\Phi(x)$, for all x, and one which also the almost sure convergences of Assumption 3.4 hold. Wherever necessary, the results in the proof are understood to hold true on this set Ω_1.

Note that

$$\frac{1}{p} \mathsf{Tr}\big(\Sigma_n^{-1} \widetilde{S}_n\big) = \frac{1}{p} \sum_{i=1}^{p} \big(u_i' \Sigma_n^{-1} u_i\big) \lambda_i\, \widetilde{\psi}_n(\lambda_i) = \int x\, \widetilde{\psi}_n(x)\, d\Phi_n(x) \;. \tag{A.1}$$

Since $\widetilde{\psi}$ is continuous and Φ_n converges weakly to Φ,

$$\int_a^b x\, \widetilde{\psi}(x)\, d\Phi_n(x) \longrightarrow \int_a^b x\, \widetilde{\psi}(x)\, d\Phi(x) \;. \tag{A.2}$$

Since $\big|\widetilde{\psi}\big|$ is continuous on $[a,b]$, it is bounded above by a finite constant $\widetilde{K}_1$. Fix $\varepsilon > 0$. Since Φ is continuous, there exists $\eta_1 > 0$ such that

$$\big|\Phi(a + \eta_1) - \Phi(a)\big| + \big|\Phi(b) - \Phi(b - \eta_1)\big| \leq \frac{\varepsilon}{6\, b\, \widetilde{K}_1} \;. \tag{A.3}$$

Since $\Phi_n(x) \longrightarrow \Phi(x)$, for all $x \in \mathbb{R}$, there exists $N_1 \in \mathbb{N}$ such that

$$\forall n \geq N_1 \qquad \max_{x \in \{a, a+\eta_1, b-\eta_1, b\}} \left|\Phi_n(x) - \Phi(x)\right| \leq \frac{\varepsilon}{24\, b\, \widetilde{K_1}} \tag{A.4}$$

Putting equations (A.3)–(A.4) together yields

$$\forall n \geq N_1 \qquad \left|\Phi_n(a+\eta_1) - \Phi_n(a)\right| + \left|\Phi_n(b) - \Phi_n(b-\eta_1)\right| \leq \frac{\varepsilon}{3\, b\, \widetilde{K_1}} \tag{A.5}$$

Therefore, for all $n \geq N_1$,

$$\begin{aligned} &\left| \int_{a+\eta_1}^{b-\eta_1} x\, \widetilde{\psi}(x)\, d\Phi_n(x) - \int_a^b x\, \widetilde{\psi}(x)\, d\Phi_n(x) \right| \\ &\qquad \leq b\widetilde{K}_1 \Big[\left|\Phi_n(a+\eta_1) - \Phi_n(a)\right| + \left|\Phi_n(b) - \Phi_n(b-\eta_1)\right| \Big] \\ &\qquad \leq \frac{\varepsilon}{3} \end{aligned} \tag{A.6}$$

Since $\widetilde{\psi}_n(x) \longrightarrow \widetilde{\psi}(x)$ uniformly over $x \in [a+\eta_1, b-\eta_1]$, there exists $N_2 \in \mathbb{N}$ such that

$$\forall n \geq N_2 \qquad \forall x \in [a+\eta_1, b-\eta_1] \qquad \left|\widetilde{\psi}_n(x) - \widetilde{\psi}(x)\right| \leq \frac{\varepsilon\, \underline{h}}{3\, b}$$

By Assumption 3.2, there exists $N_3 \in \mathbb{N}$ such that, for all $n \geq N_3$, $\max_{x \in \mathbb{R}} |\Phi_n(x)| = \mathsf{Tr}(\Sigma_n^{-1})/p$ is bounded by $1/\underline{h}$. Therefore for all $n \geq \max(N_2, N_3)$

$$\left| \int_{a+\eta_1}^{b-\eta_1} x\, \widetilde{\psi}_n(x)\, d\Phi_n(x) - \int_{a+\eta_1}^{b-\eta_1} x\, \widetilde{\psi}(x)\, d\Phi_n(x) \right| \leq b \times \frac{\varepsilon\, \underline{h}}{3\, b} \times \frac{1}{\underline{h}} = \frac{\varepsilon}{3} \tag{A.7}$$

Arguments analogous to those justifying equations (A.3)–(A.5) show there exists $N_4 \in \mathbb{N}$ such that

$$\forall n \geq N_4 \qquad \left|\Phi_n(a+\eta_1) - \Phi_n(a-\eta_1)\right| + \left|\Phi_n(b+\eta_1) - \Phi_n(b-\eta_1)\right| \leq \frac{\varepsilon}{3\, b\, \widetilde{K}} \ ,$$

for the finite constant $\widetilde{K}$ of Assumption 3.4 Therefore, for all $n \geq N_4$,

$$\left| \int_{a-\eta_1}^{b+\eta_1} x\, \widetilde{\psi}_n(x)\, d\Phi_n(x) - \int_{a+\eta_1}^{b-\eta_1} x\, \widetilde{\psi}_n(x)\, d\Phi_n(x) \right| \leq \frac{\varepsilon}{3} \tag{A.8}$$

Putting together equations (A.6)–(A.8) implies that, for all $n \geq \max(N_1, N_2, N_3, N_4)$,

$$\left| \int_{a-\eta_1}^{b+\eta_1} x\, \widetilde{\psi}_n(x)\, d\Phi_n(x) - \int_a^b x\, \widetilde{\psi}(x)\, d\Phi_n(x) \right| \leq \varepsilon$$

Since ε can be chosen aribtrarily small,

$$\int_{a-\eta_1}^{b+\eta_1} x\, \widetilde{\psi}_n(x)\, d\Phi_n(x) - \int_a^b x\, \widetilde{\psi}(x)\, d\Phi_n(x) \longrightarrow 0 \ .$$

By using equation (A.2) we get

$$\int_{a-\eta_1}^{b+\eta_1} x\, \widetilde{\psi}_n(x)\, d\Phi_n(x) \longrightarrow \int_a^b x\, \widetilde{\psi}(x)\, d\Phi(x) \ .$$

Theorem 1.1 of Bai and Silverstein (1998) shows that one a set Ω_2 of probability one, there are no sample eigenvalues outside the interval $[a-\eta_1, a+\eta_1]$, for all n large enough. Therefore, on the set $\Omega := \Omega_1 \cap \Omega_2$ of probability one,

$$\int x\,\widetilde{\psi}_n(x)\,d\Phi_n(x) \longrightarrow \int_a^b x\,\widetilde{\psi}(x)\,d\Phi(x)\ .$$

Together with equation (A.1), this proves Lemma A.2. ■

Lemma A.3.

$$\frac{1}{p}\log\Big[\det\big(\Sigma_n^{-1}\widetilde{S}_n\big)\Big] \xrightarrow{\text{a.s.}} \frac{c-1}{c}\log(1-c) - 1 + \int_a^b \log\big[\widetilde{\psi}(x)\big]\,dF(x)\ .$$

Proof of Lemma A.3.

$$\begin{aligned}
\frac{1}{p}\log\Big[\det\big(\Sigma_n^{-1}\widetilde{S}_n\big)\Big] &= \frac{1}{p}\log\Big[\det\big(\Sigma_n^{-1}\big)\det\big(\widetilde{S}_n\big)\Big] \\
&= \frac{1}{p}\log\Big[\det\big(\Sigma_n^{-1}\big)\prod_{i=1}^{p}\big(\lambda_i\widetilde{\psi}_n(\lambda_i)\big)\Big] \\
&= \frac{1}{p}\log\Big[\det\big(\Sigma_n^{-1}\big)\det\big(S_n\big)\prod_{i=1}^{p}\widetilde{\psi}_n(\lambda_i)\Big] \\
&= \frac{1}{p}\log\left[\det\left(\Sigma_n^{-1}\frac{1}{n}\sqrt{\Sigma_n}X_n'X_N\sqrt{\Sigma_n}\right)\prod_{i=1}^{p}\widetilde{\psi}_n(\lambda_i)\right] \\
&= \frac{1}{p}\log\left[\det\left(\frac{1}{n}X_n'X_n\right)\right] + \int\log\Big[\widetilde{\psi}_n(x)\Big]\,dF_n(x) \qquad \text{(A.9)}
\end{aligned}$$

Equation (1.1) of Bai and Silverstein (2004) shows that the first term on the right-hand side of equation (A.9) converges almost surely to $\big(1-c^{-1}\big)\log(1-c)-1$. As for the second term, a reasoning analogous to that conducted in the proof of Lemma A.2 shows that it converges almost surely to $\int_a^b \log\big[\widetilde{\psi}(x)\big]dF(x)$. Then Lemma A.3 follows. ■

We are now ready to tackle Theorem 3.1. Lemma A.1 and Lemma A.2 imply that

$$\frac{1}{p}\mathsf{Tr}\big(\Sigma_n^{-1}\widetilde{S}_n\big) \xrightarrow{\text{a.s.}} \int_a^b \widetilde{\psi}(x)\left\{1-c-2\,c\,x\,\mathsf{Re}[\breve{m}_F(x)]\right\}dF(x)\ .$$

Lemma A.3 implies that

$$-\frac{1}{p}\log\Big[\det\big(\Sigma_n^{-1}\widetilde{S}_n\big)\Big] - 1 \xrightarrow{\text{a.s.}} \frac{1-c}{c}\log(1-c) - \int_a^b \log\big[\widetilde{\psi}(x)\big]\,dF(x)\ .$$

Putting these two results together completes the proof of Theorem 3.1. ■

B Proof of Proposition 3.1

We start with the simpler case where $\forall n \in \mathbb{N}$, $\forall x \in \mathbb{R}$, $\widetilde{\psi}_n(x) \equiv \widetilde{\psi}(x)$. We make implicitly use of Theorem 1.1 of Bai and Silverstein (1998), which states that, for any fixed $\eta > 0$, with probability one there are no eigenvalues outside the interval $[a-\eta, b+\eta]$, for all n large enough.

For any given estimator $\widetilde{S}_n$ with limiting shrinkage function $\widetilde{\psi}$, define the bivariate function

$$\forall x, y \in [a,b] \qquad \widetilde{\psi}^\sharp(x,y) \coloneqq \begin{cases} \dfrac{x\widetilde{\psi}(x) - y\widetilde{\psi}(y)}{x-y} & \text{if} \quad x \neq y \\ x\widetilde{\psi}'(x) + \widetilde{\psi}(x) & \text{if} \quad x = y \ . \end{cases}$$

Since $\widetilde{\psi}$ is continuously differentiable on $[a,b]$, $\widetilde{\psi}^\sharp$ is continuous on $[a,b] \times [a,b]$. Consequently, there exists $K > 0$ such that, $\forall x, y \in [a,b]$, $|\widetilde{\psi}^\sharp(x,y)| \leq K$.

Lemma B.1.

$$\frac{2}{p^2} \sum_{j=1}^{p} \sum_{i>j} \frac{\lambda_j \widetilde{\psi}(\lambda_j) - \lambda_i \widetilde{\psi}(\lambda_i)}{\lambda_j - \lambda_i} \xrightarrow{\text{a.s.}} \int_a^b \int_a^b \widetilde{\psi}^\sharp(x,y)\, dF(x)\, dF(y) \ . \tag{B.1}$$

Proof of Lemma B.1.

$$\begin{aligned} \frac{2}{p^2} \sum_{j=1}^{p} \sum_{i>j} \frac{\lambda_j \widetilde{\psi}(\lambda_j) - \lambda_i \widetilde{\psi}(\lambda_i)}{\lambda_j - \lambda_i} &= \frac{1}{p^2} \sum_{j=1}^{p} \sum_{i=1}^{p} \widetilde{\psi}^\sharp(\lambda_i, \lambda_j) - \frac{1}{p^2} \sum_{j=1}^{p} \widetilde{\psi}^\sharp(\lambda_j, \lambda_j) \\ &= \int_a^b \int_a^b \widetilde{\psi}^\sharp(x,y)\, dF_n(x)\, dF_n(y) - \frac{1}{p^2} \sum_{j=1}^{p} \widetilde{\psi}^\sharp(\lambda_j, \lambda_j) \ . \end{aligned}$$

Given equation (3.1), the first term converges almost surely to the right-hand side of equation (B.1). The absolute value of the second term is bounded by K/p; therefore, it vanishes asymptotically. ■

Lemma B.2.

$$\int_a^b \int_a^b \widetilde{\psi}^\sharp(x,y)\, dF(x)\, dF(y) = -2 \int_a^b x\widetilde{\psi}(x)\, \mathsf{Re}\left[\breve{m}_F(x)\right]\, dF(x) \tag{B.2}$$

Proof of Lemma B.2. Fix any $\varepsilon > 0$. Then there exists $\eta_1 > 0$ such that, for all $v \in (0, \eta_1)$,

$$\left| 2 \int_a^b x\widetilde{\psi}(x)\, \mathsf{Re}\left[\breve{m}_F(x)\right]\, dF(x) - 2 \int_a^b x\widetilde{\psi}(x)\, \mathsf{Re}\left[\breve{m}_F(x+iv)\right]\, dF(x) \right| \leq \frac{\varepsilon}{4} \ .$$

The definition of the Stieltjes transform implies

$$-2 \int_a^b x\widetilde{\psi}(x)\, \mathsf{Re}\left[\breve{m}_F(x+iv)\right]\, dF(x) = 2 \int_a^b \int_a^b \frac{x\widetilde{\psi}(x)(x-y)}{(x-y)^2 + v^2}\, dF(x)\, dF(y) \ .$$

There exists $\eta_2 > 0$ such that, for all $v \in (0, \eta_1)$,

$$\left| 2 \int_a^b \int_a^b \frac{x\widetilde{\psi}(x)(x-y)}{(x-y)^2+v^2} dF(x) dF(y) - 2 \int_a^b \int_a^b \frac{x\widetilde{\psi}(x)(x-y)}{(x-y)^2+v^2} \mathbb{1}_{\{|x-y| \geq \eta_2\}} dF(x) dF(y) \right| \leq \frac{\varepsilon}{4}$$

$$\text{and} \qquad \left| \int_a^b \int_a^b \widetilde{\psi}^\sharp(x,y)\, dF(x)\, dF(y) - \int_a^b \int_a^b \widetilde{\psi}^\sharp(x,y) \mathbb{1}_{\{|x-y| \geq \eta_2\}}\, dF(x)\, dF(y) \right| \leq \frac{\varepsilon}{4} \ .$$

We have

$$\begin{aligned}\int_a^b \int_a^b \widetilde{\psi}^\sharp(x,y)\mathbb{1}_{\{|x-y|\geq\eta_2\}}\,dF(x)\,dF(y) &= \int_a^b \int_a^b \frac{x\widetilde{\psi}(x)-y\widetilde{\psi}(y)}{x-y}\mathbb{1}_{\{|x-y|\geq\eta_2\}}\,dF(x)\,dF(y)\\ &= \int_a^b \int_a^b \frac{x\widetilde{\psi}(x)}{x-y}\mathbb{1}_{\{|x-y|\geq\eta_2\}}\,dF(x)\,dF(y)\\ &\quad + \int_a^b \int_a^b \frac{y\widetilde{\psi}(y)}{y-x}\mathbb{1}_{\{|y-x|\geq\eta_2\}}\,dF(y)\,dF(x)\\ &= 2\int_a^b \int_a^b \frac{x\widetilde{\psi}(x)}{x-y}\mathbb{1}_{\{|x-y|\geq\eta_2\}}\,dF(x)\,dF(y)\end{aligned}$$

Note that

$$\begin{aligned}2\int_a^b \int_a^b \frac{x\widetilde{\psi}(x)}{x-y}\mathbb{1}_{\{|x-y|\geq\eta_2\}}\,dF(x)\,dF(y) - 2\int_a^b \int_a^b \frac{x\widetilde{\psi}(x)(x-y)}{(x-y)^2+v^2}\mathbb{1}_{\{|x-y|\geq\eta_2\}}\,dF(x)\,dF(y)\\ = 2\int_a^b \int_a^b \frac{x\widetilde{\psi}(x)}{x-y}\frac{v^2}{(x-y)^2+v^2}\mathbb{1}_{\{|x-y|\geq\eta_2\}}\,dF(x)\,dF(y)\ ,\end{aligned}$$

and that

$$\forall(x,y) \text{ such that } |x-y|\geq\eta_2 \qquad \frac{v^2}{(x-y)^2+v^2} \leq \frac{v^2}{\eta_2^2+v^2}\ .$$

The quantity on the right-hand side can be made arbitrarily small for fixed η_2 by bringing v sufficiently close to zero. This implies that there exists $\eta_3 \in (0,\eta_1)$ such that, for all $v \in (0,\eta_3)$,

$$\left|2\int_a^b \int_a^b \frac{x\widetilde{\psi}(x)}{x-y}\mathbb{1}_{\{|x-y|\geq\eta_2\}}\,dF(x)\,dF(y) - 2\int_a^b \int_a^b \frac{x\widetilde{\psi}(x)(x-y)}{(x-y)^2+v^2}\mathbb{1}_{\{|x-y|\geq\eta_2\}}\,dF(x)\,dF(y)\right| \leq \frac{\varepsilon}{4}\,.$$

Putting these results together yields

$$\left|\int_a^b \int_a^b \widetilde{\psi}^\sharp(x,y)\,dF(x)\,dF(y) + 2\int_a^b x\widetilde{\psi}(x)\,\mathsf{Re}\left[\breve{m}_F(x)\right]\,dF(x)\right| \leq \varepsilon\ .$$

Since this holds for any $\varepsilon > 0$, equation (B.2) follows. ■

Putting together Lemmas B.1 and B.2 yields

$$\frac{2}{p^2}\sum_{j=1}^{p}\sum_{i>j}\frac{\lambda_j\widetilde{\psi}(\lambda_j)-\lambda_i\widetilde{\psi}(\lambda_i)}{\lambda_j-\lambda_i} \xrightarrow{\text{a.s.}} -2\int_a^b x\widetilde{\psi}(x)\,\mathsf{Re}\left[\breve{m}_F(x)\right]\,dF(x)\ .$$

Lemma B.3. *As n and p go to infinity with their ratio p/n converging to the concentration c,*

$$\log(n) - \frac{1}{p}\sum_{j=1}^{p}\mathbb{E}[\log(\chi^2_{n-j+1})] \longrightarrow 1 + \frac{1-c}{c}\log(1-c)\ .$$

Proof of Lemma B.3. It is well known that, for every positive integer ν,

$$\mathbb{E}[\log(\chi^2_\nu)] = \log(2) + \frac{\Gamma'(\nu/2)}{\Gamma(\nu/2)}\ ,$$

where $\Gamma(\cdot)$ denotes the gamma function. Thus

$$\frac{1}{p}\sum_{j=1}^{p}\mathbb{E}[\log(\chi^2_{n-j+1})] = \log(2) + \frac{1}{p}\sum_{j=1}^{p}\frac{\Gamma'((n-j+1)/2)}{\Gamma((n-j+1)/2)} \ .$$

Formula 6.3.21 of Abramowitz and Stegun (1965) states that

$$\forall x \in (0,+\infty) \qquad \frac{\Gamma'(x)}{\Gamma(x)} = \log(x) - \frac{1}{2x} - 2\int_0^\infty \frac{t\,dt}{(t^2+x^2)(e^{2\pi t}-1)} \ .$$

It implies that

$$\begin{aligned}\log(n) - \frac{1}{p}\sum_{j=1}^{p}\mathbb{E}[\log(\chi^2_{n-j+1})] &= -\frac{1}{p}\sum_{j=1}^{p}\log\left(1-\frac{j-1}{n}\right) + \frac{1}{p}\sum_{k=n-p+1}^{n}\frac{1}{k} \\ &\quad + \frac{1}{p}\sum_{k=n-p+1}^{n}\int_0^\infty \frac{t\,dt}{[t^2+(k/2)^2](e^{2\pi t}-1)} \\ &=: -\frac{1}{p}\sum_{j=1}^{p}\log\left(1-\frac{j-1}{n}\right) + A_n + B_n \ .\end{aligned}$$

It is easy to verify that

$$-\frac{1}{p}\sum_{j=1}^{p}\log\left(1-\frac{j-1}{n}\right) \longrightarrow -\frac{1}{c}\int_0^c \log(1-x)dx = 1 + \frac{1-c}{c}\log(1-c) \ .$$

Therefore, all that remains to be proven is that the two terms A_n and B_n vanish. Using Formulæ 6.3.2 and 6.3.18 of Abramowitz and Stegun (1965), we see that

$$A_n := \frac{1}{p}\sum_{k=n-p+1}^{n}\frac{1}{k} = \frac{1}{p}\left[\frac{\Gamma'(n)}{\Gamma(n)} - \frac{\Gamma'(n-p+1)}{\Gamma(n-p+1)}\right] = \frac{1}{p}\log\left(\frac{n}{n-p+1}\right) + O\left(\frac{1}{p(n-p+1)}\right) \ ,$$

which vanishes indeed. As for the term B_n, it admits the upper bound

$$B_n := \frac{1}{p}\sum_{k=n-p+1}^{n}\int_0^\infty \frac{t\,dt}{[t^2+(k/2)^2](e^{2\pi t}-1)} \leq \int_0^\infty \frac{t\,dt}{[t^2+((n-p+1)/2)^2](e^{2\pi t}-1)} \ ,$$

which also vanishes. ■

Going back to equation (2.1), we notice that the term

$$\frac{2}{p}\sum_{j=1}^{p}\lambda_j\widetilde{\psi}'(\lambda_j)$$

remains bounded asymptotically with probability one, since $\widetilde{\psi}'$ is bounded over a compact set.

Putting all these results together shows that the unbiased estimator of risk $\Theta_n(S_n, \widehat{\Sigma})$ converges almost surely to

$$(1-c)\int_a^b \widetilde{\psi}(x)dF(x) - \int_a^b \log[\widetilde{\psi}(x)]dF(x) - 2c\int_a^b x\widetilde{\psi}(x)\mathsf{Re}[\breve{m}_F(x)]dF(x) + \frac{1-c}{c}\log(1-c) \ .$$

It is easy to verify that these results carry through to the more general case where the shrinkage function $\widetilde{\psi}_n$ can vary across n, as long as it is well behaved asymptotically in the sense of Assumption 3.4. ■

C Proof of Proposition 3.2

We provide a proof by contradiction. Suppose that Proposition 3.2 does not hold. Then there exist $\varepsilon > 0$ and $x_0 \in \mathsf{Supp}(F)$ such that

$$1 - c - 2\,c\,x_0\,\mathsf{Re}[\breve{m}_F(x_0)] \leq \frac{a_1}{\overline{h}} - 2\varepsilon \ . \tag{C.1}$$

Since $\breve{m}_F$ is continuous, there exist $x_1, x_2 \in \mathsf{Supp}(F)$ such that $x_1 < x_2$, $[x_1, x_2] \subset \mathsf{Supp}(F)$, and

$$\forall x \in [x_1, x_2] \qquad 1 - c - 2\,c\,x\,\mathsf{Re}[\breve{m}_F(x)] \leq \frac{a_1}{\overline{h}} - \varepsilon \ .$$

Define, for all $n \in \mathbb{N}$ and $x \in \mathbb{R}$,

$$\begin{aligned}
\overline{\psi}(x) &\coloneqq \mathbb{1}_{[x_1,x_2]}(x) \\
\overline{\psi}_n(x) &\coloneqq \overline{\psi}(x) \\
\overline{D}_n &\coloneqq \mathsf{Diag}\left(\lambda_{n,1}\overline{\psi}_n(\lambda_{n,1}), \ldots, \lambda_{n,p}\overline{\psi}_n(\lambda_{n,p})\right) \\
\overline{S}_n &\coloneqq U_n \overline{D}_n U_n' \ .
\end{aligned}$$

By Lemmas A.1–A.2,

$$\frac{1}{p}\mathsf{Tr}\left(\Sigma_n^{-1}\overline{S}_n\right) \xrightarrow{\text{a.s.}} \int \overline{\psi}(x)\left\{1 - c - 2\,c\,x\,\mathsf{Re}[\breve{m}_F(x)]\right\} dF(x) \ . \tag{C.2}$$

The left-hand side of equation (C.2) is asymptotically bounded from below as follows.

$$\begin{aligned}
\frac{1}{p}\mathsf{Tr}\left(\Sigma_n^{-1}\overline{S}_n\right) &= \frac{1}{p}\sum_{i=1}^{p} u_{n,i}'\Sigma_n^{-1}u_{n,i} \times \lambda_{n,i}\,\mathbb{1}_{[x_1,x_2]}(\lambda_{n,i}) \\
&\geq \frac{\lambda_{n,1}}{\overline{h}}\left[F_n(x_2) - F_n(x_1)\right] \xrightarrow{\text{a.s.}} \frac{a_1}{\overline{h}}\left[F(x_2) - F(x_1)\right] \ .
\end{aligned} \tag{C.3}$$

The right-hand side of equation (C.2) is bounded from above as follows.

$$\int \overline{\psi}(x)\left\{1 - c - 2\,c\,x\,\mathsf{Re}[\breve{m}_F(x)]\right\} dF(x) \leq \left(\frac{a_1}{\overline{h}} - \varepsilon\right)\left[F(x_2) - F(x_1)\right] \ . \tag{C.4}$$

Given that $F(x_2) - F(x_1) > 0$, equations (C.2)–(C.4) form a logical contradiction. Therefore, the initial assumption (C.1) must be false, which proves Proposition 3.2. ■

D Proof of Theorem 4.2

Define the shrinkage function

$$\forall x \in \mathsf{Supp}\big(F_{n,p}^{\widehat{\boldsymbol{\tau}}_n}\big) \qquad \widehat{\psi}_n(x) \coloneqq \frac{1}{1 - \dfrac{p}{n} - 2\,\dfrac{p}{n}\,x\,\mathsf{Re}\big[\breve{m}_{n,p}^{\widehat{\boldsymbol{\tau}}_n}(x)\big]} \ .$$

Theorem 2.2 of Ledoit and Wolf (2013) and Proposition 4.3 of Ledoit and Wolf (2012) imply that $\forall x \in \mathsf{Supp}(F)$ $\widehat{\psi}_n(x) \xrightarrow{\text{a.s.}} \psi^*(x)$, and that this convergence is uniform over $x \in \mathsf{Supp}(F)$, apart from arbitrarily small boundary regions of the support. Theorem 4.2 then follows from Corollary 3.1. ■