Ledoit, Olivier; Wolf, Michael

Working Paper

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

Working Paper, No. 122 [rev.]

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 [rev.], University of Zurich, Department of Economics, Zurich, https://doi.org/10.5167/uzh-78074

This Version is available at:
https://hdl.handle.net/10419/92396

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

Revised version, December 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

First version: May 2013
This version: December 2013

Abstract

This paper introduces a new method for deriving covariance matrix estimators that are decision-theoretically optimal. The key is to employ large-dimensional asymptotics: the matrix dimension and the sample size go to infinity together, with their ratio converging to a finite, nonzero limit. As the main focus, we apply this method to Stein's loss. Compared to the estimator of Stein (1975, 1986), ours has five theoretical advantages:

1. it asymptotically minimizes the loss itself, instead of an *estimator* of the expected loss;
2. it does not necessitate post-processing through an *ad hoc* algorithm (called "isotonization") to restore the positivity or the ordering of the covariance matrix eigenvalues;
3. it does not ignore any terms in the function to be minimized;
4. it does not require normality; and
5. it is not limited to applications where the sample size exceeds the dimension.

In addition to these theoretical advantages, our estimator also improves upon Stein' estimator in terms of finite-sample performance, as evidenced via extensive Monte Carlo simulations.

To further demonstrate the effectiveness of our method, we show that some previously suggested estimators of the covariance matrix and its inverse are decision-theoretically optimal with respect to the Frobenius loss function.

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

JEL CLASSIFICATION NOS: C13.

*Olivier Ledoit wishes to thank Stanford University Statistics Department seminar participants, and especially Bala Rajaratnam, for feedback on an earlier version of this paper.

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, signal processing, neuroimaging, climatology, 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 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 articles by 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 Stein's estimator is still considered the *gold standard* (Rajaratnam et al., 2013) and has proven hard to beat empirically, a careful reading of Stein's original articles reveals several theoretical limitations.

1. The estimator proposed by Stein (1975, 1986) does not minimize the loss, nor the risk (that is, the expected loss), but instead an unbiased estimator of the risk. This is problematic because the primary objects of interest are the loss and the risk. *A priori* there could exist many unbiased estimators of the risk, so that minimizing them could lead to different estimators. Furthermore, the resulting estimators may not minimize the primary objects of interest: the loss or the risk.
2. The formula derived by Stein generates covariance matrix estimators that may not be positive semidefinite. 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 in the unbiased estimator of the risk that involves the derivatives of the shrinkage function. No justification, apart from tractability, is given for this omission.
4. Stein's estimator requires normality, an assumption often violated by real data.
5. Finally, Stein's estimator is only defined when the sample size exceeds the dimension.

One important reason why Stein's estimator is highly regarded in spite of its 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* isotonizing algorithm.

Our paper develops a shrinkage estimator of the covariance matrix in the spirit of Stein (1975, 1986) with two significant improvements: first, it solves the five 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 his class of rotation-equivariant estimators that have the same eigenvectors as the sample covariance matrix.

The key difference is that we 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 acknowledges the usefulness of RMT. But he uses it for illustration purposes only, which opens up the question of whether RMT could contribute more than that and deliver an improved Stein-type estimator of the covariance matrix. Important new results in RMT enable us to answer these questions positively in the present paper.

We show that, under a certain set of assumptions, 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, the corresponding risk, and Stein's unbiased estimator of the risk are all asymptotically equivalent;
2. bypass the need for an isotonizing algorithm;
3. justify that the term involving the derivatives of the shrinkage function (which was ignored by Stein) vanishes indeed;
4. dispense with the normality assumption; and
5. handle the challenging case where the dimension exceeds the sample size.

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

Our procedure is divided into two distinct steps: first, we find an *oracle* estimator that

is asymptotically optimal but depends on unobservable population quantities; second, we find a *bona fide* estimator that depends only on observable quantities, is asymptotically equivalent to the oracle, and thus inherits its the oracle's asymptotic optimality property. The second step is not original, as we adapt technology developed earlier by Ledoit and Wolf (2012, 2013). However, the first step is a key original contribution of the present paper, made possible by the introduction of the new concept of *limiting shrinkage function*. In order to demonstrate its effectiveness, we apply it to the estimators of Ledoit and Wolf (2012, 2013) and prove that these previously suggested estimators are asymptotically optimal with respect to their respective loss functions. (This optimality result strengthens the two earlier papers.) In passing, we unearth deep, unexpected connections between Stein's loss and the quadratic loss functions used by Ledoit and Wolf (2012, 2013).

Additional evidence for our method being effective is the fact that it enables us to discover a new oracle covariance matrix estimator which is optimal with respect to the *Symmetrized Stein's loss*. Not only does this estimator aim to be close to the population covariance matrix, but at the same time it aims to have an inverse close to the inverse of the population covariance matrix. Such symmetry is mathematically elegant and points to a promising avenue for future research.

The remainder of the paper is organized as follows. Section 2 briefly summarizes the finite-sample theory of Stein (1975, 1986). Section 3 details the adjustments necessary to transplant Stein's theory from finite samples to large-dimensional asymptotics. Section 4 showcases the effectiveness of our new method for deriving oracle estimators of the covariance matrix that are asymptotically optimal with respect to various loss functions. Section 5 develops our feasible estimator of a covariance matrix, which is asymptotically optimal with respect to Stein's loss. Section 6 extends the analysis to the challenging case where the matrix dimension exceeds the sample size, the sample covariance matrix is singular, and Stein's estimator is not even defined. Section 7 studies finite-sample properties via Monte Carlo simulations. Section 8 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

$$\forall x \in \mathbb{R} \qquad H_n(x) \coloneqq \frac{1}{p} \sum_{i=1}^{p} \mathbb{1}_{[\tau_{n,i}, +\infty)}(x) \ ,$$

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 $\sqrt{\ }$ denotes the symmetric positive-definite square root of the matrix. 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.

Definition 2.1 (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: $\widetilde{D}_n \coloneqq (\widetilde{\varphi}_n(\lambda_{n,1}) \ldots, \widetilde{\varphi}_n(\lambda_{n,p}))$, and $\widetilde{\varphi}_n$ 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 estimate of the covariance matrix. 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 $\widetilde{\varphi}_n$ 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 $\widetilde{\varphi}_n(\cdot)$ itself but with the vector $(\widetilde{\varphi}_{n,1}, \ldots, \widetilde{\varphi}_{n,p})' \coloneqq \big(\widetilde{\varphi}_n(\lambda_{n,1}), \ldots, \widetilde{\varphi}_n(\lambda_{n,p})\big)'$ instead. This is equivalent because the sample eigenvalues are distinct with probability one, and because the values taken by the shrinkage function $\widetilde{\varphi}_n(\cdot)$ 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 e.d.f. of the sample eigenvalues,

$$\forall x \in \mathbb{R} \qquad F_n(x) \coloneqq \frac{1}{p} \sum_{i=1}^{p} \mathbb{1}_{[\lambda_{n,i}, +\infty)}(x) \ ,$$

than with the vector of the sample eigenvalues.

Definition 2.2 (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:*

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

and its corresponding risk function $\mathcal{R}_n^S(\Sigma_n, \widetilde{S}_n) \coloneqq \mathbb{E}[\mathcal{L}_n^S(\Sigma_n, \widetilde{S}_n)]$*. Here, we introduce* $\mathsf{Tr}(\cdot)$ *as the notation for the trace operator.*

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 $\mathcal{R}_n^S(\Sigma_n, \widetilde{S}_n)$ as *Stein's risk*.

2.2 Stein's Loss in Finite Samples

Stein (1986) introduces a function closely related to the nonlinear shrinkage function: $\widetilde{\psi}(x) \coloneqq \widetilde{\varphi}(x)/x$. Under Assumptions 2.1–2.3, Stein shows that the risk function satisfies the identity $\mathcal{R}_n^S(\Sigma_n, \widetilde{S}_n) = \mathbb{E}[\Theta_n(\Sigma_n, \widetilde{S}_n)]$, 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 $\widetilde{\varphi}_n$ satisfies $\forall i = 1, \ldots, p,\ \ \widetilde{\varphi}_n(\lambda_{n,i}) = \varphi_n^*(\lambda_{n,i})$, where

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

Although 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 therefore necessitates post-processing through an *ad hoc* isotonizing algorithm whose effect is hard to quantify theoretically; for example, see the insightful work of Rajaratnam et al. (2013). Eschewing isotonization is one of our 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 φ_n^{ST}, Stein's *isotonized* estimator is

$$S_n^{ST} := U_n D_n^{ST} U_n' \ , \quad \text{where} \quad D_n^{ST} := \mathsf{Diag}\big(\varphi_n^{ST}(\lambda_{n,1}), \ldots, \varphi_n^{ST}(\lambda_{n,p})\big) \ . \tag{2.3}$$

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, where the dimension goes to infinity together with the sample size.

3.1 Large-Dimensional Asymptotic Framework

Assumption 3.1 (Dimension)**.** *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 $c \in (0,1)$ called the* limiting concentration*. Furthermore, there exists a compact interval included in $(0,1)$ that contains p/n for all n large enough.*

The extension to the case $p > n$ is covered in Section 6.

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 $\boldsymbol{\tau}_n := (\tau_{n,1}, \ldots, \tau_{n,p})'$ 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 both standard assumptions in the literature on large-dimensional asymptotics; see Bai and Silverstein (2010) for a comprehensive review. Though less widespread, the assumption that $\mathsf{Supp}(H_n)$ is uniformly bounded away from infinity has been made by such authors as Bai and Silverstein (1998, 1999, 2004) and El Karoui (2008), among others. It precludes, for example, the factor model where all pairs of variables have 50% correlation and all variables have unit standard deviation. However, we would argue that Stein's class of rotation-equivariant estimators and his rotation-invariant loss function are ill-suited in this case. The reason is that a factor model of this type displays such strong orientation: flipping the sign of half the variables (which is a rotation) is not indifferent because it destroys the structure of the problem. Having said that, Monte Carlo simulations reported in Figure 5 indicate that our estimator performs well in practice even when the largest eigenvalue goes to infinity.

The population eigenvalues are uniformly bounded away from zero in the widely used 'spiked' model of Johnstone (2001), as well as in the papers published by Mestre (2008), Ledoit and Péché (2011), Ledoit and Wolf (2012), and Won et al. (2012). (Note that all these papers also require the largest eigenvalue to be bounded away from infinity.) In theory, we

estimate well-conditioned covariance matrices. In practice, Monte Carlo simulations reported in Figure 2 indicate that our estimator performs well even when the lower bound of the support of H approaches zero.

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 $Y_n \coloneqq X_n \times \sqrt{\Sigma_n}$. Neither $\sqrt{\Sigma_n}$ nor X_n are observed on their own: only Y_n is observed.*

Note that we no longer require normality.

Remark 3.1 (Moment condition)**.** The existence of a finite 12th moment is assumed to prove certain mathematical results using the methodology of Ledoit and Péché (2011). However, Monte Carlo studies in Ledoit and Wolf (2012, 2013) indicate that this assumption is not needed in practice and can be replaced with the existence of a finite fourth moment. ■

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

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

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. Although 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 the sample eigenvalues. Assumptions 3.1–3.3 together with Bai and Silverstein (1998, Theorem 1.1) imply that the support of F, denoted by $\mathsf{Supp}(F)$, 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 $0 < a_1 < b_1 < \cdots < a_\kappa < b_\kappa < \infty$.

Definition 3.1 (Estimators)**.** *We consider covariance matrix estimators of the type* $\widetilde{S}_n := U_n \widetilde{D}_n U_n'$*, where* $\widetilde{D}_n$ *is a diagonal matrix:* $\widetilde{D}_n := \mathsf{Diag}\big(\widetilde{\varphi}_n(\lambda_{n,i}) \ldots, \widetilde{\varphi}_n(\lambda_{n,i})\big)$ *and* $\widetilde{\varphi}_n$ *is a (possibly random) real univariate function which may depend on* S_n*.*

Assumption 3.4. *We assume that there exists a nonrandom real univariate function* $\widetilde{\varphi}$ *defined on* $\mathsf{Supp}(F)$ *and continuously differentiable on* $\bigcup_{k=1}^{\kappa}[a_k, b_k]$ *such that* $\widetilde{\varphi}_n(x) \xrightarrow{\text{a.s.}} \widetilde{\varphi}(x)$ *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* $x \in \bigcup_{k=1}^{\kappa}[a_k - \eta, b_k + \eta]$*,* $|\widetilde{\varphi}_n(x)|$ *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 $\widetilde{\varphi}$.

Definition 3.2 (Loss Function)**.** *Estimators are evaluated according to the limit, as* n *and* p *go to infinity together, of the following loss function:*

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

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

The key difference is that, instead of minimizing the unbiased estimator of risk $\Theta_n(\Sigma_n, \widetilde{S}_n)$ defined in equation (2.1), as Stein (1975, 1986) does, we minimize $\lim_{p,n\to_c\infty}\mathcal{L}_n^S(\Sigma_n, \widetilde{S}_n)$ and $\lim_{p,n\to_c\infty}\Theta_n(\Sigma_n, \widetilde{S}_n)$. Here, we introduce the notation "$p, n \to_c \infty$" as indicating that both p and n go to infinity together, with their ratio p/n converging to a constant c; see Assumption 3.1.

The almost sure existence and equality of these two limits is established below.

3.2 Stein's Loss under Large-Dimensional Asymptotics

Theorem 3.1. *Under Assumptions 3.1–3.4,*

$$\mathcal{L}_n^S(\Sigma_n, \widetilde{S}_n) \xrightarrow{\text{a.s.}} \sum_{k=1}^{\kappa}\int_{a_k}^{b_k}\left\{\frac{1 - c - 2\,c\,x\,\mathsf{Re}[\breve{m}_F(x)]}{x}\,\widetilde{\varphi}(x) - \log[\widetilde{\varphi}(x)]\right\}dF(x)$$
$$+ \int \log(t)\,dH(t) - 1\ . \tag{3.5}$$

The proof is in Appendix A.1. 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 3.1–3.4,*

$$\Theta_n(\Sigma_n, \widetilde{S}_n) \xrightarrow{\text{a.s.}} \sum_{k=1}^{\kappa} \int_{a_k}^{b_k} \left\{ \frac{1 - c - 2\,c\,x\,\mathsf{Re}[\breve{m}_F(x)]}{x}\, \widetilde{\varphi}(x) - \log[\widetilde{\varphi}(x)] \right\} dF(x) + \int \log(t)\, dH(t) - 1 \ . \tag{3.6}$$

The proof is in Appendix A.2. Proposition 3.1 shows that, under large-dimensional asymptotics, minimizing the unbiased estimator of risk is asymptotically 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 by Stein 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. *Suppose Assumptions 3.1–3.4 hold.*

(i) A covariance matrix estimator $\widetilde{S}_n$ minimizes in the class of rotation-equivariant estimators described in Definition 3.1 the almost sure limit (3.5) *of Stein's loss if and only if its limiting shrinkage function $\widetilde{\varphi}$ verifies $\forall x \in \mathsf{Supp}(F)$, $\widetilde{\varphi}(x) = \varphi^*(x)$, where*

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

The resulting oracle estimator of the covariance matrix is

$$S_n^* \coloneqq U_n \times \mathsf{Diag}\big(\varphi^*(\lambda_{n,1}), \ldots, \varphi^*(\lambda_{n,p})\big) \times U_n' \ .$$

(ii) The minimum of the almost sure limit (3.5) *of Stein's loss is equal to*

$$\lim_{p,n \to_c \infty} \mathcal{L}_n^S(\Sigma_n, S_n^*) = \int \log(t)\, dH(t) - \sum_{k=1}^{\kappa} \int_{a_k}^{b_k} \log\left[\frac{x}{1 - c - 2\,c\,x\,\mathsf{Re}[\breve{m}_F(x)]} \right] dF(x) \ . \tag{3.8}$$

Equation (3.7) follows immediately from Theorem 3.1 by differentiating the right-hand side of equation (3.5) with respect to $\widetilde{\varphi}(x)$. Equation (3.8) obtains by plugging equation (3.7) into equation (3.5) and simplifying.

The fact that the denominator on the right-hand side of equation (3.7) is nonzero and that the optimal limiting shrinkage function φ^* is strictly positive and bounded over the support of F is established by the following proposition, whose proof is in Appendix A.3.

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 covariance matrix estimator based on the nonlinear shrinkage function φ^* is an *oracle* estimator, as it depends on m_F, the Stieltjes transform of the limiting spectral distribution of the sample covariance matrix. m_F is unobservable, as it depends on H, the limiting spectral distribution of the population covariance matrix, which is itself unobservable. Nonetheless, as we will show in Section 5, this oracle estimator plays a pivotal role because it is the foundation on which a *bona fide* estimator enjoying the same asymptotic optimality properties can be erected.

4 Beyond Stein's Loss

Although the present paper focuses mainly on Stein's loss and the nonlinear shrinkage function φ^*, a key innovation relative to Ledoit and Wolf (2012, 2013) is the method of Section 3.2 for finding an oracle estimator that minimizes the limit of a prespecified loss function under large-dimensional asymptotics; or, alternatively, for proving that an existing estimator is asymptotically optimal with respect to some specific loss function. It is important to demonstrate that the effectiveness of this method extends beyond Stein's loss. Since Section 4 constitutes a digression from the central theme of the paper as stated in the title itself, we limit ourselves to loss functions that either are intimately related to Stein's loss or have been previously used by Ledoit and Wolf (2012, 2013).

4.1 Inverse Stein's Loss

The first natural extension is to apply Stein's loss to the inverse of the covariance matrix, also called the *precision matrix*. Equation (1.3) of Tsukuma (2005) thus defines the loss function

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

Its limit is given by the following theorem, whose proof is in Appendix B.1.

Theorem 4.1. *Under Assumptions 3.1–3.4,*

$$\begin{aligned}\mathcal{L}_n^{SINV}(\Sigma_n, \widetilde{S}_n) \xrightarrow{\text{a.s.}} \sum_{k=1}^{\kappa} \int_{a_k}^{b_k} \left\{ \frac{x}{\left|1 - c - c\,x\,\breve{m}_F(x)\right|^2\ \widetilde{\varphi}(x)} + \log[\widetilde{\varphi}(x)] \right\} dF(x) \\ - \int \log(t)\, dH(t) - 1\ . \end{aligned} \tag{4.1}$$

Differentiating the right-hand side of equation (4.1) with respect to $\widetilde{\varphi}(x)$ yields an oracle estimator that is optimal with respect to the Inverse Stein's loss in large dimensions.

Corollary 4.1. *Under Assumptions 3.1–3.4, a covariance matrix estimator $\widetilde{S}_n$ minimizes in the class of rotation-equivariant estimators described in Definition 3.1 the almost sure limit of the Inverse Stein's loss if and only if its limiting shrinkage function $\widetilde{\varphi}$ verifies $\forall x \in \mathsf{Supp}(F)$, $\widetilde{\varphi}(x) = \varphi^\circ(x)$, where*

$$\forall x \in \mathsf{Supp}(F) \qquad \varphi^\circ(x) \coloneqq \frac{x}{\left|1 - c - c\,x\,\breve{m}_F(x)\right|^2}\ . \tag{4.2}$$

4.2 Frobenius Loss

Ledoit and Wolf (2012, 2013) use the following loss function based on the squared Frobenius distance:

$$\mathcal{L}_n^{F}(\Sigma_n, \widetilde{S}_n) \coloneqq \frac{1}{p}\mathsf{Tr}\left[\left(\Sigma_n - \widetilde{S}_n\right)^2\right]\ .$$

Its limit is given by the following theorem, whose proof is in Appendix B.2.

Theorem 4.2. *Under Assumptions 3.1–3.4,*

$$\mathcal{L}_n^F(\Sigma_n, \widetilde{S}_n) \xrightarrow{\text{a.s.}} \int x^2 \, dH(x) + \sum_{k=1}^{\kappa} \left\{ -2 \int_{a_k}^{b_k} \frac{x \, \widetilde{\varphi}(x)}{|1 - c - c\, x \, \breve{m}_F(x)|^2} dF(x) + \int_{a_k}^{b_k} \widetilde{\varphi}(x)^2 dF(x) \right\}. \tag{4.3}$$

Differentiating the right-hand side of equation (4.3) with respect to $\widetilde{\varphi}(x)$ enables us to characterize the set of asymptotically optimal estimators under the Frobenius loss in large dimensions.

Corollary 4.2. *Under Assumptions 3.1–3.4, a covariance matrix estimator $\widetilde{S}_n$ minimizes in the class of rotation-equivariant estimators described in Definition 3.1 the almost sure limit of the Frobenius loss if and only if its limiting shrinkage function $\widetilde{\varphi}$ verifies $\forall x \in \mathsf{Supp}(F)$, $\widetilde{\varphi}(x) = \varphi^{\circ}(x)$.*

To the best of our knowledge, the close relationship between Frobenius loss and Inverse Stein's loss had not been observed before.

Both Ledoit and Wolf (2012, Section 3.1) and Ledoit and Wolf (2013, Section 3) use the Frobenius loss and the oracle nonlinear shrinkage estimator φ°. But in these two papers the justification for using this oracle estimator is different (namely, as an approximation to the *finite-sample optimal* estimator). Therefore, Corollary 4.2 strengthens these two earlier papers by providing a more formal justification for the oracle estimator they use.

4.3 Inverse Frobenius Loss

Ledoit and Wolf (2012, Section 3.2) apply the Frobenius loss to the precision matrix:

$$\mathcal{L}_n^{FINV}(\Sigma_n, \widetilde{S}_n) \coloneqq \mathcal{L}_n^F(\Sigma_n^{-1}, \widetilde{S}_n^{-1}) = \frac{1}{p} \mathsf{Tr} \left[\left(\Sigma_n^{-1} - \widetilde{S}_n^{-1} \right)^2 \right] .$$

Its limit is given by the following theorem, whose proof is in Appendix B.3.

Theorem 4.3. *Under Assumptions 3.1–3.4,*

$$\mathcal{L}_n^{FINV}(\Sigma_n, \widetilde{S}_n) \xrightarrow{\text{a.s.}} \int \frac{dH(x)}{x^2} + \sum_{k=1}^{\kappa} \left\{ -2 \int_{a_k}^{b_k} \frac{1 - c - 2cx\mathsf{Re}[\breve{m}_F(x)]}{x \, \widetilde{\varphi}(x)} dF(x) + \int_{a_k}^{b_k} \frac{1}{\widetilde{\varphi}(x)^2} dF(x) \right\}. \tag{4.4}$$

Differentiating the right-hand side of equation (4.4) with respect to $\widetilde{\varphi}(x)$ enables us to characterize the set of asymptotically optimal estimators under the Inverse Frobenius loss in large dimensions.

Corollary 4.3. *Under Assumptions 3.1–3.4, a covariance matrix estimator $\widetilde{S}_n$ minimizes in the class of rotation-equivariant estimators described in Definition 3.1 the almost sure limit of the Inverse Frobenius loss if and only if its limiting shrinkage function $\widetilde{\varphi}$ verifies $\forall x \in \mathsf{Supp}(F)$, $\widetilde{\varphi}(x) = \varphi^{*}(x)$.*

The Inverse Frobenius loss yields the same oracle estimator as Stein's loss. This surprising mathematical result shows that a *bona fide* covariance matrix estimator based on the nonlinear shrinkage function φ^{*} — which we shall obtain in Section 5 — can be justified in multiple ways.

4.4 Symmetrized Stein's Loss

The correspondence between Stein's loss and Frobenius loss is crossed. The shrinkage function φ^* should be used to estimate the *covariance* matrix according to Stein's loss, and to estimate the *precision* matrix according to Frobenius loss. According to Stein's loss, the function φ° optimally estimates the precision matrix, but according to Frobenius loss, it optimally estimates the covariance matrix instead. Thus, if we are interested in estimating the covariance matrix, but have no strong preference between Stein's loss and Frobenius loss, should we take φ^* or φ°? Similarly, if a researcher needs a good estimator of the precision matrix, but has no opinion on the relative merits of Stein's loss versus Frobenius loss, should we recommend φ° or φ^*?

In the machine learning literature, loss functions that pay equal attention to the twin problems of estimating the covariance matrix and estimating its inverse take pride of place. A representative example is equation (17.8) of Moakher and Batchelor (2006).[1] The *Symmetrized* Stein's loss (function) is defined as

$$\mathcal{L}_n^{SSYM}(\Sigma_n, \widetilde{S}_n) := \frac{\mathcal{L}_n^S(\Sigma_n, \widetilde{S}_n) + \mathcal{L}_n^S(\Sigma_n^{-1}, \widetilde{S}_n^{-1})}{2} = \frac{1}{2p}\mathsf{Tr}\left(\Sigma_n^{-1}\widetilde{S}_n + \Sigma_n\widetilde{S}_n^{-1}\right) - 1 \ .$$

This loss function is symmetric in the sense that $\mathcal{L}_n^{SSYM}(\Sigma_n, \widetilde{S}_n) = \mathcal{L}_n^{SSYM}(\Sigma_n^{-1}, \widetilde{S}_n^{-1})$, and also in the sense that $\mathcal{L}_n^{SSYM}(\Sigma_n, \widetilde{S}_n) = \mathcal{L}_n^{SSYM}(\widetilde{S}_n, \Sigma_n)$. Its limit is given by the following theorem.

Theorem 4.4. *Under Assumptions 3.1–3.4,*

$$\begin{aligned}\mathcal{L}_n^{SSYM}(\Sigma_n, \widetilde{S}_n) \xrightarrow{\text{a.s.}} & \frac{1}{2}\sum_{k=1}^{\kappa}\int_{a_k}^{b_k}\frac{1 - c - 2\,c\,x\,\mathsf{Re}[\breve{m}_F(x)]}{x}\,\widetilde{\varphi}(x)dF(x) \\ & + \frac{1}{2}\sum_{k=1}^{\kappa}\int_{a_k}^{b_k}\frac{x}{|1 - c - c\,x\,\breve{m}_F(x)|^2\ \widetilde{\varphi}(x)}\,dF(x) - 1 \ . \end{aligned} \tag{4.5}$$

The proof follows trivially from Theorems 3.1 and 4.1 and is thus omitted. Differentiating the right-hand side of equation (4.5) with respect to $\widetilde{\varphi}(x)$ enables us to characterize the set of asymptotically optimal estimators under the Symmetrized Stein's loss in large dimensions.

Corollary 4.4. *Under Assumptions 3.1–3.4, a covariance matrix estimator $\widetilde{S}_n$ minimizes in the class of rotation-equivariant estimators described in Definition 3.1 the almost sure limit of the Symmetrized Stein's loss if and only if its limiting shrinkage function $\widetilde{\varphi}$ verifies $\forall x \in \mathsf{Supp}(F)$, $\widetilde{\varphi}(x) = \varphi^{\circledast}(x)$, where*

$$\forall x \in \mathsf{Supp}(F) \qquad \varphi^{\circledast}(x) := \sqrt{\varphi^*(x)\varphi^\circ(x)} \ . \tag{4.6}$$

This nonlinear shrinkage function has not been discovered before. The resulting oracle estimator of the covariance matrix is $S_n^{\circledast} := U_n \times \mathsf{Diag}\big(\varphi^{\circledast}(\lambda_{n,1}), \ldots, \varphi^{\circledast}(\lambda_{n,p})\big) \times U_n'$. This estimator is generally attractive because it strikes a balance between the covariance matrix and its inverse, and also between Stein's loss and Frobenius loss. Furthermore, Jensen's inequality guarantees that $\forall x \in \mathbb{R},\ \varphi^*(x) < \varphi^{\circledast}(x) < \varphi^\circ(x)$.

[1] We thank an anonymous referee for bringing this reference to our attention.

4.5 Synthesis

Section 4 constitutes somewhat of a digression from the central theme of the paper, but we can take away from it several important points:

1. Given that a key technical innovation of the present paper is the method for obtaining oracle estimators that are asymptotically optimal with respect to some prespecified loss function, Section 4 demonstrates that this method can handle a variety of loss functions.

2. This method also strengthens the earlier papers of Ledoit and Wolf (2012, 2013) by providing a more formal justification for their oracle estimators.

3. The oracle estimator that is optimal with respect to Stein's loss turns out to be also optimal with respect to the Inverse Frobenius loss, an unexpected connection. Conversely, the oracle estimator that is optimal with respect to the Inverse Stein's loss is also optimal with respect to the Frobenius loss.

4. The covariance matrix estimator that is optimal with respect to the Symmetrized Stein's loss is both new and interesting in that it is equally attentive to both the covariance matrix *and* its inverse. Modern analyses such as Moakher and Batchelor's (2006) indicate that this is a desirable property for loss functions defined on the Riemannian manifold of symmetric positive-definite matrices. To wit, Stein's loss does not even define a proper notion of distance, whereas Stein's Symmetrized loss does.

5 Optimal Covariance Matrix Estimation

The procedure for going from an oracle estimator to a *bona fide* estimator has been developed by Ledoit and Wolf (2012, 2013). Here we repeat it for convenience, adapting it to Stein's loss. The basic idea is to first obtain a consistent estimator of the eigenvalues of the population covariance matrix and to then derive from it a consistent estimator of the Stieltjes transform of the limiting sample spectral distribution.

5.1 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

$$Q_{n,p} : [0,\infty)^p \longrightarrow [0,\infty)^p \tag{5.1}$$

$$\mathbf{t} \coloneqq (t_1,\ldots,t_p)' \longmapsto Q_{n,p}(\mathbf{t}) \coloneqq (q^1_{n,p}(\mathbf{t}),\ldots,q^p_{n,p}(\mathbf{t}))' \;, \tag{5.2}$$

where

$$\forall i = 1, \ldots, p \qquad q_{n,p}^{i}(\mathbf{t}) \coloneqq p \int_{(i-1)/p}^{i/p} \left(F_{n,p}^{\mathbf{t}}\right)^{-1}(u)\, du \ , \tag{5.3}$$

$$\forall u \in [0,1] \qquad \left(F_{n,p}^{\mathbf{t}}\right)^{-1}(u) \coloneqq \sup\{x \in \mathbb{R} : F_{n,p}^{\mathbf{t}}(x) \leq u\} \ , \tag{5.4}$$

$$\forall x \in \mathbb{R} \qquad F_{n,p}^{\mathbf{t}}(x) \coloneqq \lim_{\eta \to 0^+} \frac{1}{\pi} \int_{-\infty}^{x} \mathsf{Im}\left[m_{n,p}^{\mathbf{t}}(\xi + i\eta)\right] d\xi \ , \tag{5.5}$$

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

$$\left\{ m \in \mathbb{C} : -\frac{n-p}{nz} + \frac{p}{n}\, m \in \mathbb{C}^+ \right\} \tag{5.6}$$

to the equation

$$m = \frac{1}{p} \sum_{i=1}^{p} \frac{1}{t_i \left(1 - \dfrac{p}{n} - \dfrac{p}{n}\, z\, m\right) - z} \ . \tag{5.7}$$

It can be seen that equation (5.5) quantizes equation (3.4), that equation (5.6) quantizes equation (3.2), and that equation (5.7) 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 (5.4), $\left(F_{n,p}^{\mathbf{t}}\right)^{-1}$ represents the inverse spectral distribution function, also known as the *quantile* function. By equation (5.3), $q_{n,p}^{i}(\mathbf{t})$ can be interpreted as a 'smoothed' version of the $(i - 0.5)/p$ quantile of $F_{n,p}^{\mathbf{t}}$.

5.2 Consistent Estimator of the Population Eigenvalues

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

Theorem 5.1. *Suppose that Assumptions 3.1–3.3 are satisfied. Define*

$$\widehat{\boldsymbol{\tau}}_n \coloneqq \underset{\mathbf{t} \in (0,\infty)^p}{\operatorname{argmin}} \frac{1}{p} \sum_{i=1}^{p} \left[q_{n,p}^{i}(\mathbf{t}) - \lambda_{n,i}\right]^2 , \tag{5.8}$$

where $\boldsymbol{\lambda}_n \coloneqq (\lambda_{n,1}, \ldots, \lambda_{n,p})'$ *are the eigenvalues of the sample covariance matrix* S_n*, and* $Q_{n,p}(\mathbf{t}) \coloneqq (q_{n,p}^{1}(\mathbf{t}), \ldots, q_{n,p}^{p}(\mathbf{t}))'$ *is the nonrandom QuEST function defined in equations* (5.1)–(5.7)*; 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*

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

The proof is given by Ledoit and Wolf (2013, Theorem 2.2). The solution to equation (5.8) can be found by standard nonlinear optimization software such as SNOPT$^{\text{TM}}$; see Gill et al. (2002).

5.3 Asymptotically Optimal Estimator of the Covariance Matrix

Recall that, for any $\mathbf{t} \coloneqq (t_1, \ldots, t_p)' \in (0, +\infty)^p$, equations (5.6)–(5.7) 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 $\forall \lambda \in \mathbb{R}$, $\lim_{z\in\mathbb{C}^+\to\lambda} m_{n,p}^{\mathbf{t}}(z) =: \breve{m}_{n,p}^{\mathbf{t}}(\lambda)$ 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 (5.8). The proof of Theorem 5.2 is in Appendix C.1.

Theorem 5.2. *Under Assumptions 3.1–3.4, the covariance matrix estimator*

$$\begin{aligned} &\widehat{S}_n^* \coloneqq U_n \widehat{D}_n^* U_n' \quad \textit{where} \quad \widehat{D}_n^* \coloneqq \mathsf{Diag}\big(\widehat{\varphi}_n^*(\lambda_{n,1}), \ldots, \widehat{\varphi}_n^*(\lambda_{n,p})\big) \\ &\textit{and} \quad \forall i = 1, \ldots, p \quad \widehat{\varphi}_n^*(\lambda_{n,i}) \coloneqq \frac{\lambda_{n,i}}{1 - \dfrac{p}{n} - 2\,\dfrac{p}{n}\,\lambda_{n,i}\,\mathsf{Re}\big[\breve{m}_{n,p}^{\widehat{\boldsymbol{\tau}}_n}(\lambda_{n,i})\big]} \end{aligned} \tag{5.9}$$

minimizes in the class of rotation-equivariant estimators described in Definition 3.1 the almost sure limit (3.5) *of Stein's loss as n and p go to infinity together.*

Remark 5.1 (Alternative loss functions)**.** Similarly, plugging the consistent estimator $\breve{m}_{n,p}^{\widehat{\boldsymbol{\tau}}_n}$ in place of the unobservable $\breve{m}_F$ in the oracle estimators derived in Section 4 yields *bona fide* covariance matrix estimators that minimize the almost sure limits of their respective loss functions. In the case of Inverse Stein's loss and Frobenius loss, the resulting optimal estimator $\widehat{S}^{\circ}$ is the same as the estimator defined by Ledoit and Wolf (2013). In the case of Inverse Frobenius loss, the resulting optimal estimator is $\widehat{S}^*$. In the case of Symmetrized Stein's loss, the resulting optimal estimator is $\widehat{S}^{\circledast} \coloneqq \sqrt{\widehat{S}^*\widehat{S}^{\circ}}$. A further study of the estimator $\widehat{S}^{\circledast}$, involving a comprehensive set of Monte Carlo simulations to examine finite-sample performance, lies beyond the scope of the present paper and is left for future research. ■

Both Stein (1975) and the present paper attack the same problem with two very different mathematical techniques, so how far apart are the resulting estimators? The answer hinges on the concept of *Cauchy Principal Value* (PV). The convolution of a compactly supported function $g(t)$ with the Cauchy kernel $(t-x)^{-1}$ is generally an improper integral due to the singularity at $t = x$. However there is a way to properly define this convolution as

$$\forall x \in \mathbb{R} \qquad G(x) \coloneqq \mathrm{PV}\int_{-\infty}^{\infty} \frac{g(t)}{t-x}\,dt \coloneqq \lim_{\varepsilon \searrow 0}\left[\int_{-\infty}^{x-\varepsilon} \frac{g(t)}{t-x}\,dt + \int_{x+\varepsilon}^{\infty} \frac{g(t)}{t-x}\,dt\right].$$

Henrici (1988, pp. 259–262) is a useful reference for Principal Values. Stein's shrinkage function and ours (equations (2.2) and (5.9) respectively) can be expressed as

$$\forall i = 1, \ldots, p \qquad \varphi_n^*(\lambda_{n,i}) \coloneqq \frac{\lambda_{n,i}}{1 - \dfrac{p-1}{n} - 2\,\dfrac{p}{n}\,\lambda_{n,i} \times \mathrm{PV}\displaystyle\int_{-\infty}^{\infty} \frac{1}{\lambda - \lambda_{n,i}}\,dF_n(\lambda)}$$

$$\forall i = 1, \ldots, p \qquad \widehat{\varphi}_n^*(\lambda_{n,i}) \coloneqq \frac{\lambda_{n,i}}{1 - \dfrac{p}{n} - 2\,\dfrac{p}{n}\,\lambda_{n,i} \times \mathrm{PV}\displaystyle\int_{-\infty}^{\infty} \frac{1}{\lambda - \lambda_{n,i}}\,dF_{n,p}^{\widehat{\boldsymbol{\tau}}_n}(\lambda)}$$

The only material difference is that the step function F_n is replaced by the smooth function $F_{n,p}^{\widehat{\tau}_n}$. It is reassuring that two approaches using such unrelated mathematical techniques generate concordant results.

Both F_n and $F_{n,p}^{\widehat{\tau}_n}$ estimate the limiting sample spectral distribution F, but not in the same way: the former is the "naïve" estimator, while the latter is the product of cutting-edge research in Random Matrix Theory. Convolving the Cauchy kernel with a step function such as F_n is dangerously unstable when two consecutive steps happen to be too close to each other. This is why Stein's original estimator needs to be regularized *ex post* through the isotonizing algorithm. By contrast, our estimator of the sample spectral distribution is sufficiently regular *ex ante* to admit convolution with the Cauchy kernel without creating instability. This is why our approach is more elegant in theory, and also has the potential to be more accurate in practice, as Monte Carlo simulations in Section 7 will confirm.

6 Extension to the Singular Case

So far, we have only considered the case $p < n$, as does Stein (1975, 1986). We do not know whether Stein was not interested in the case singular case $p > n$ or whether he could not solve the problem of how to then shrink the zero eigenvalues of the sample covariance matrix. Either way, another key contribution of the present paper is that we can also handle this challenging case. Assumption 3.1 now has to be modified as follows.

Assumption 6.1 (Dimension)**.** *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 $c \in (1, \infty)$ called the* limiting concentration*. Furthermore, there exists a compact interval included in $(1, \infty)$ that contains p/n for all n large enough.*

Under Assumption 6.1, F is a mixture distribution with mass $(c-1)/c$ at zero and a continuous component whose compact support is bounded away from zero; for example, see Ledoit and Wolf (2013, Section 2.1). Define

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

so that $\underline{F}$ corresponds to the continuous component of F, normalized to be a proper distribution (function).

Now Assumptions 3.2, 3.3, and 6.1 together with Bai and Silverstein (1998, Theorem 1.1) imply that the support of $\underline{F}$, denoted by $\mathsf{Supp}(\underline{F})$, is the union of a finite number $\kappa \geq 1$ of compact intervals: $\mathsf{Supp}(\underline{F}) = \bigcup_{k=1}^{\kappa}[a_k, b_k]$, where $0 < a_1 < b_1 < \cdots < a_\kappa < b_\kappa < \infty$. Furthermore, $\mathsf{Supp}(F) = \{0\} \cup \mathsf{Supp}(\underline{F})$. Note that with this notation, there is no further need to modify Assumption 3.4.

As a first step in deriving the *bona fide* estimator, we establish the almost sure existence of the limit of Stein's loss in the case $p > n$.

Theorem 6.1. *Under Assumptions 3.2–3.4 and 6.1,*

$$\mathcal{L}_n^S(\Sigma_n, \widetilde{S}_n) \xrightarrow{\text{a.s.}} \sum_{k=1}^{\kappa} \int_{a_k}^{b_k} \left\{ \frac{1 - c - 2\,c\,x\,\mathsf{Re}[\breve{m}_F(x)]}{x}\,\widetilde{\varphi}(x) - \log[\widetilde{\varphi}(x)] \right\} dF(x) \\ + \int \log(t)\,dH(t) + \frac{c-1}{c} \left\{ \Big[\frac{c}{c-1} \cdot \breve{m}_H(0) - \breve{m}_{\underline{F}}(0)\Big]\widetilde{\varphi}(0) - \log[\widetilde{\varphi}(0)] \right\} - 1 \ . \tag{6.1}$$

The proof is in Appendix D.1. As a second step, Theorem 6.1 enables us to characterize the set of asymptotically optimal estimators under Stein's loss in large dimensions in the case $p > n$.

Corollary 6.1. *Suppose Assumptions 3.2–3.4 and 6.1 hold.*

(i) A covariance matrix estimator $\widetilde{S}_n$ minimizes in the class of rotation-equivariant estimators described in Definition 3.1 the almost sure limit (6.1) *of Stein's loss if and only if its limiting shrinkage function $\widetilde{\varphi}$ verifies $\forall x \in \mathsf{Supp}(F)$, $\widetilde{\varphi}(x) = \varphi^*(x)$, where*

$$\varphi^*(0) \coloneqq \left(\frac{c}{c-1} \cdot \breve{m}_H(0) - \breve{m}_{\underline{F}}(0) \right)^{-1}, \\ \text{and} \quad \forall x \in \mathsf{Supp}(\underline{F}) \qquad \varphi^*(x) \coloneqq \frac{x}{1 - c - 2\,c\,x\,\mathsf{Re}[\breve{m}_F(x)]} \ . \tag{6.2}$$

The resulting oracle estimator of the covariance matrix is

$$S_n^* \coloneqq U_n \times \mathsf{Diag}\big(\varphi^*(\lambda_{n,1}), \ldots, \varphi^*(\lambda_{n,p})\big) \times U_n' \ .$$

(ii) The minimum of the almost sure limit (6.1) *of Stein's loss is equal to*

$$\lim_{p,n \to_c \infty} \mathcal{L}_n^S(\Sigma_n, S_n^*) = \int \log(t)\,dH(t) - \sum_{k=1}^{\kappa} \int_{a_k}^{b_k} \log\left[\frac{x}{1 - c - 2\,c\,x\,\mathsf{Re}[\breve{m}_F(x)]} \right] dF(x) \\ + \frac{c-1}{c} \log\Big[\frac{c}{c-1} \cdot \breve{m}_H(0) - \breve{m}_{\underline{F}}(0)\Big] \ . \tag{6.3}$$

Equation (6.2) follows immediately from Theorem 6.1 by differentiating the right-hand side of equation (6.1) with respect to $\widetilde{\varphi}(x)$. Equation (6.3) obtains by plugging equation (6.2) into equation (6.1) and simplifying.

As a third step, the procedure for going from the oracle estimator to the *bona fide* estimator is similar to the case $p < n$. But we also have to find strongly consistent estimators of the quantities $\breve{m}_H(0)$ and $\breve{m}_{\underline{F}}(0)$ which did not appear in the oracle shrinkage function in the case $p < n$.

Let $\widehat{\boldsymbol{\tau}}_n \coloneqq (\widehat{\tau}_{n,1}, \ldots, \widehat{\tau}_{n,p})'$ denote the set of estimated eigenvalues defined as in Theorem 5.1. A strongly consistent estimator of $\breve{m}_H(0)$ is given by

$$\widehat{\breve{m}_H}(0) \coloneqq \frac{1}{p} \sum_{i=1}^{p} \frac{1}{\widehat{\tau}_{n,i}} \ . \tag{6.4}$$

As explained in Ledoit and Wolf (2013, Section 3.2.2), a strongly consistent estimator of the quantity $\breve{m}_{\underline{F}}(0)$ is the unique solution $m =: \widehat{\breve{m}_{\underline{F}}}(0)$ in $(0, \infty)$ to the equation

$$m = \left[\frac{1}{n} \sum_{i=1}^{p} \frac{\widehat{\tau}_{n,i}}{1 + \widehat{\tau}_{n,i}\, m} \right]^{-1} . \tag{6.5}$$

Theorem 6.2. *Under Assumptions 3.2–3.4 and 6.1, the covariance matrix estimator*

$$\widehat{S}_n^* := U_n \widehat{D}_n^* U_n' \quad \textit{where} \quad \widehat{D}_n^* := \mathsf{Diag}\big(\widehat{\varphi}_n^*(\lambda_{n,1}), \ldots, \widehat{\varphi}_n^*(\lambda_{n,p})\big) \ ,$$

$$\forall i = 1, \ldots, p-n \quad \widehat{\varphi}_n^*(\lambda_{n,i}) := \left(\frac{p/n}{p/n - 1} \cdot \widehat{\breve{m}_H}(0) - \widehat{\breve{m}_{\underline{F}}}(0) \right)^{-1} ,$$

$$\textit{and} \quad \forall i = p-n+1, \ldots, p \quad \widehat{\varphi}_n^*(\lambda_{n,i}) := \frac{\lambda_{n,i}}{1 - \frac{p}{n} - 2\,\frac{p}{n}\,\lambda_{n,i}\, \mathsf{Re}\big[\breve{m}_{n,p}^{\widehat{\boldsymbol{\tau}}_n}(\lambda_{n,i})\big]}$$

minimizes in the class of rotation-equivariant estimators described in Definition 3.1 the almost sure limit (6.1) *of Stein's loss.*

The proof is in Appendix D.2.

7 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 5.2 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

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

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

Unless stated otherwise, 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, and 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 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.

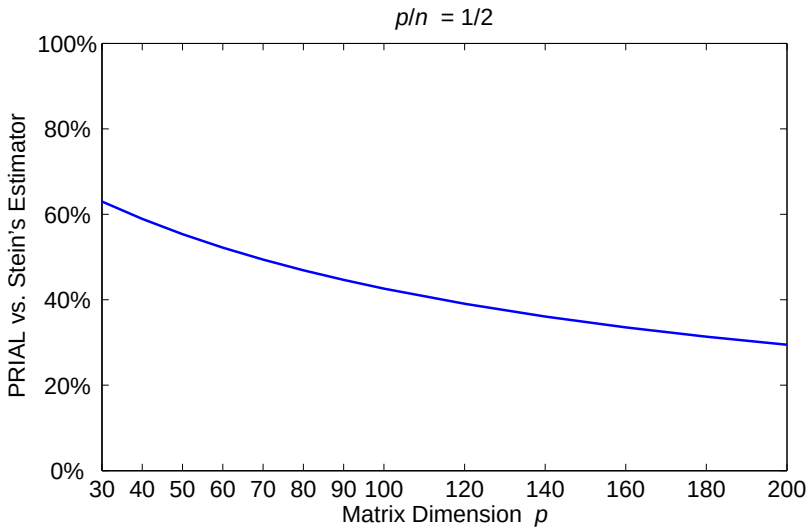

Figure 1: Evolution of the PRIAL of the new nonlinear shrinkage estimator relative to Stein's estimator as matrix dimension and sample size go to infinity together.

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.

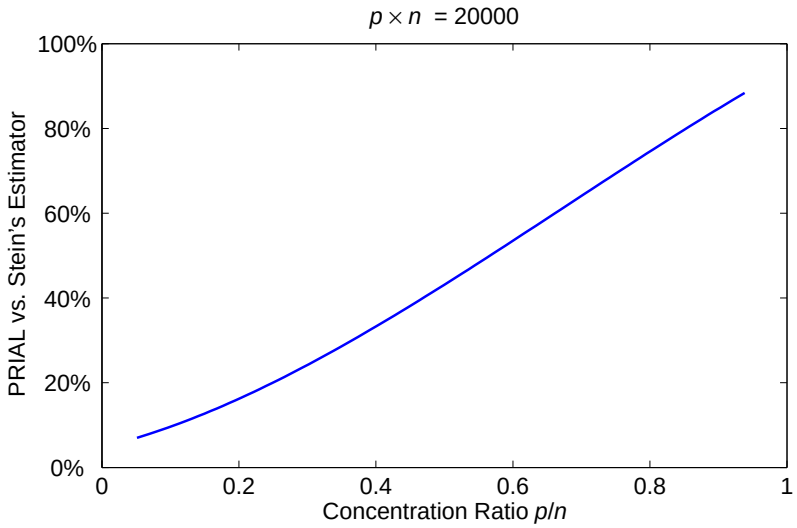

Figure 2: PRIAL of the new nonlinear shrinkage estimator relative to Stein's estimator as a function of the concentration ratio p/n.

One can see 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.

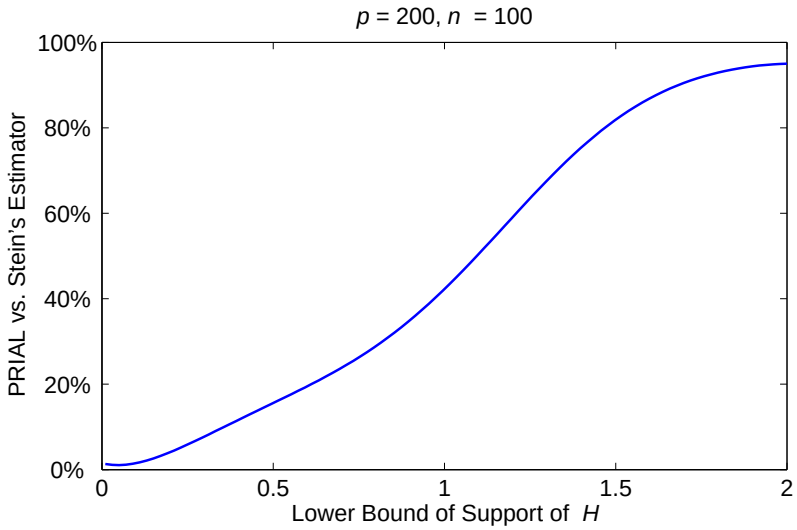

Figure 3: PRIAL of the new nonlinear shrinkage estimator relative to Stein's estimator across various condition numbers.

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.

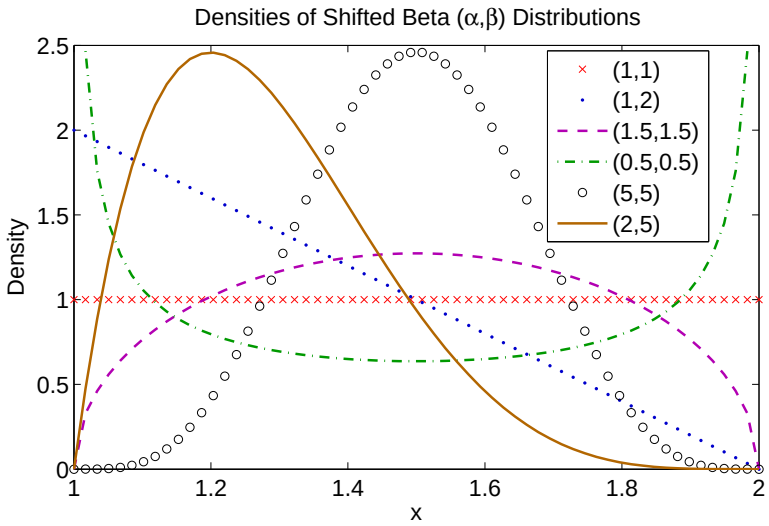

Figure 4: Densities of various shifted Beta distributions. Note that the density of Beta(β, α) is just the mirror image (around the mid point of the support) of the density of Beta(α, β).

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 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

Fifth, we vary the distribution of the variates X_n. Beyond the (standard) normal distribution with kurtosis 0, we also consider the coin-toss Bernoulli distribution, which is platykurtic with kurtosis -2, and the (standard) 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 for various distributions of the variates.

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

Unbounded Top Eigenvalue

Sixth, we study the performance of our estimator in the case where the largest population eigenvalue is of order n, in violation of Assumption 3.2. Inspired by the factor model described below Assumption 3.2, we set $\tau_{n,p}$ equal to $1 + 0.5(p-1)$. The other eigenvalues are set as per the baseline scenario. Thus, $\boldsymbol{\tau}_n \coloneqq (H^{-1}(0.5/(p-1)), \ldots, H^{-1}((p-1.5)/(p-1)), 1+0.5(p-1))'$, where H is the distribution of $1 + W$, and $W \sim \text{Beta}(2, 5)$. The dimension ranges from $p = 30$ to $p = 200$, and the concentration ratio p/n is fixed at the value 1/2. The results are displayed in Figure 5.

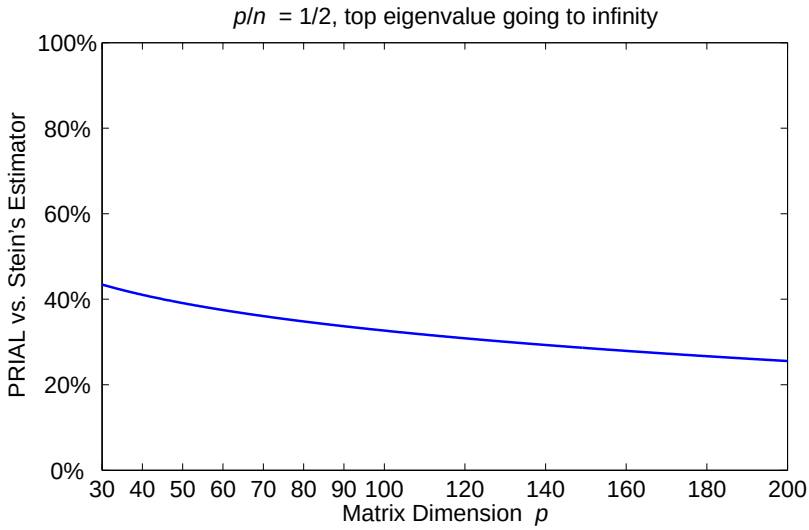

Figure 5: PRIAL of the new nonlinear shrinkage estimator relative to Stein's estimator when the top eigenvalue diverges.

Our estimator still dominates convincingly Stein's estimator, even though Assumption 3.2 is violated.

Singular Case

Finally, we study the challenging case $p > n$ where the sample covariance matrix is singular and Stein's estimator is not defined. We set the concentration ratio $c = p/n$ equal to two, take the same distribution for H as in the baseline case, and simulate Gaussian variates. The dimension ranges from $p = 30$ to $p = 400$. The benchmark is the minimum of the almost sure limit of Stein's loss in the class of nonlinear shrinkage estimators; see equation (6.3). For this choice of H and c, the minimum is equal to 0.007232385 (evaluated numerically). The average loss across $1,000$ Monte Carlo simulations for our nonlinear shrinkage estimator is displayed in Figure 6.

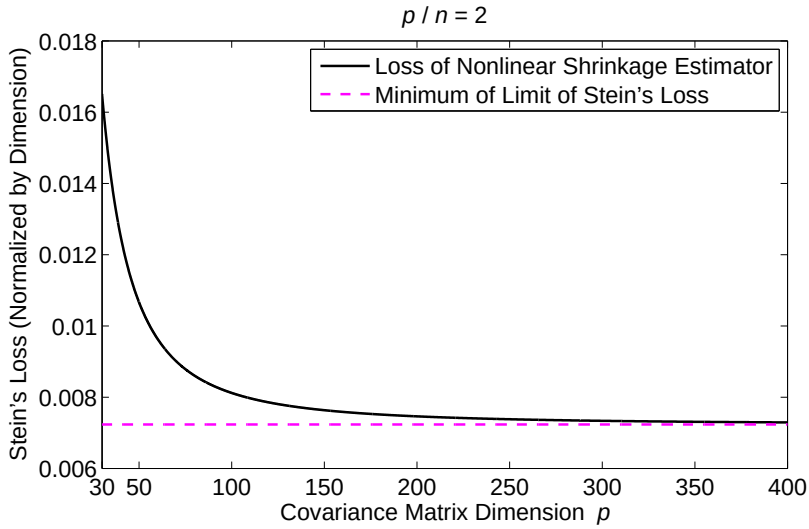

Figure 6: Stein's Loss for the nonlinear shrinkage estimator when dimension exceeds sample size. The benchmark is the minimum of the limit of Stein's loss among rotation-equivariant estimators.

These results confirm that our nonlinear shrinkage estimator minimizes Stein's loss asymptotically even in the difficult case where variables outnumber observations.

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

8 Concluding Remarks

Estimating a covariance matrix is one of the two 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 revisits the pioneering work of Stein (1975, 1986) to construct an improved estimator of a covariance matrix, based on the scale-invariant loss function commonly known as

Stein's loss. The estimator originally proposed by Stein suffers from a certain number of limitations, among which the two most visible ones are: first, the possibility of violation of eigenvalue ordering; and second, 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's estimator minimizes an unbiased estimator of risk in finite samples, within 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. Unlike Stein's estimator, ours also works when the dimension exceeds the sample size.

Our paper represents an original contribution not only with respect to Stein's papers but also with respect to the recent literature on large-dimensional asymptotics. Indeed, our asymptotic optimality results, made possible by the introduction of the new concept of *limiting shrinkage function*, provide a more formal justification to estimators based on the Frobenius loss proposed by Ledoit and Wolf (2012, 2013).

We use a two-step method, whereby we first derive an optimal oracle estimator using our new technique, and then find an equivalent *bona fide* estimator using methodology developed by Ledoit and Wolf (2012, 2013). The end product is a covariance matrix estimator that minimizes the almost sure limit of the loss function in the class of nonlinear shrinkage estimators, as sample size and dimension go to infinity together.

When applied to Stein's loss, our method delivers an estimator that both circumvents the theoretical difficulties that beset Stein's estimator and also enjoys improved finite-sample performance, as evidenced by extensive simulations.

An in-depth study of estimators that are asymptotically optimal with respect to other loss functions, such as Symmetrized Stein's loss, is beyond the scope of this paper but points to promising avenues for future research.

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. (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. (2004). CLT for linear spectral statistics of large-dimensional sample covariance matrices. *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.

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

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.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal component analysis. *Annals of Statistics*, 29(2):295–327.

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.

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.

Moakher, M. and Batchelor, P. G. (2006). Symmetric positive-definite matrices: From geometry to applications and visualization. In *Visualization and Processing of Tensor Fields*, pages 285–298. Springer.

Rajaratnam, B., Vincenzi, D., and Naul, B. (2013). A study of Stein's covariance estimator within the unbiased estimator of risk framework. *Journal of Multivariate Analysis — (in revision).*

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.

Tsukuma, H. (2005). Estimating the inverse matrix of scale parameters in an elliptically contoured distribution. *Journal of the Japan Statistical Society*, 35(1):21–39.

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 in the case $p < n$, 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. On the same grounds, we make a similar assumption on the support of $\underline{F}$ in the case $p > n$; see Section 6.

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

A Proof of Mathematical Results in Section 3.2

A.1 Proof of Theorem 3.1

Definition A.1. *For any integer k, define $\forall x \in \mathbb{R}$, $\Delta_n^{(k)}(x) \coloneqq p^{-1} \sum_{i=1}^{p} u_i' \Sigma_n^k u_i \times \mathbb{1}_{[\lambda_i, +\infty)}(x)$.*

Lemma A.1. *Under Assumptions 3.1–3.3, there exists a nonrandom function $\Delta^{(-1)}$ defined on $\mathbb{R}$ such that $\Delta_n^{(-1)}(x)$ converges almost surely to $\Delta^{(-1)}(x)$, for all $x \in \mathbb{R}$. Furthermore, $\Delta^{(-1)}$ is continuously differentiable on $\mathbb{R}$ and satisfies $\forall x \in \mathbb{R}$, $\Delta^{(-1)}(x) = \int_{-\infty}^{x} \delta^{(-1)}(\lambda) dF(\lambda)$, where*

$$\forall \lambda \in \mathbb{R} \qquad \delta^{(-1)}(\lambda) \coloneqq \begin{cases} 0 & \text{if } \lambda \leq 0, \\ \dfrac{1 - c - 2\, c\, \lambda\, \mathsf{Re}[\breve{m}_F(\lambda)]}{\lambda} & \text{if } \lambda > 0 \ . \end{cases}$$

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,*

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

Proof of Lemma A.2. Restrict attention to the set Ω_1 of probability one on which $\Delta_n^{(-1)}(x)$ converges to $\Delta^{(-1)}(x)$, for all x, and on 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

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

Since $\widetilde{\varphi}$ is continuous and $\Delta_n^{(-1)}$ converges weakly to $\Delta^{(-1)}$,

$$\int_a^b \widetilde{\varphi}(x)\, d\Delta_n^{(-1)}(x) \longrightarrow \int_a^b \widetilde{\varphi}(x)\, d\Delta^{(-1)}(x) \ . \tag{A.2}$$

Since $\left|\widetilde{\varphi}\right|$ is continuous on $[a,b]$, it is bounded above by a finite constant $\widetilde{K}_1$. Fix $\varepsilon > 0$. Since $\Delta^{(-1)}$ is continuous, there exists $\eta_1 > 0$ such that

$$\left|\Delta^{(-1)}(a+\eta_1) - \Delta^{(-1)}(a)\right| + \left|\Delta^{(-1)}(b) - \Delta^{(-1)}(b-\eta_1)\right| \leq \frac{\varepsilon}{6\,\widetilde{K}_1} \ . \tag{A.3}$$

Since $\Delta_n^{(-1)}(x) \longrightarrow \Delta^{(-1)}(x)$, for all $x \in \mathbb{R}$, there exists $N_1 \in \mathbb{N}$ such that

$$\forall n \geq N_1 \qquad \max_{x \in \{a, a+\eta_1, b-\eta_1, b\}} \left|\Delta_n^{(-1)}(x) - \Delta^{(-1)}(x)\right| \leq \frac{\varepsilon}{24\,\widetilde{K}_1} \ . \tag{A.4}$$

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

$$\forall n \geq N_1 \qquad \left|\Delta_n^{(-1)}(a+\eta_1) - \Delta_n^{(-1)}(a)\right| + \left|\Delta_n^{(-1)}(b) - \Delta_n^{(-1)}(b-\eta_1)\right| \leq \frac{\varepsilon}{3\,\widetilde{K}_1} \ . \tag{A.5}$$

Therefore, for all $n \geq N_1$,

$$\begin{aligned}
&\left| \int_{a+\eta_1}^{b-\eta_1} \widetilde{\varphi}(x)\, d\Delta_n^{(-1)}(x) - \int_a^b \widetilde{\varphi}(x)\, d\Delta_n^{(-1)}(x) \right| \\
&\qquad\qquad \leq \widetilde{K}_1 \Big[\left|\Delta_n^{(-1)}(a+\eta_1) - \Delta_n^{(-1)}(a)\right| + \left|\Delta_n^{(-1)}(b) - \Delta_n^{(-1)}(b-\eta_1)\right| \Big] \\
&\qquad\qquad \leq \frac{\varepsilon}{3} \ .
\end{aligned} \tag{A.6}$$

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

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

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

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

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

$$\forall n \geq N_4 \qquad \left|\Delta_n^{(-1)}(a+\eta_1) - \Delta_n^{(-1)}(a-\eta_1)\right| + \left|\Delta_n^{(-1)}(b+\eta_1) - \Delta_n^{(-1)}(b-\eta_1)\right| \leq \frac{\varepsilon}{3\,\widetilde{K}} \ ,$$

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

$$\left| \int_{a-\eta_1}^{b+\eta_1} \widetilde{\varphi}_n(x)\, d\Delta_n^{(-1)}(x) - \int_{a+\eta_1}^{b-\eta_1} \widetilde{\varphi}_n(x)\, d\Delta_n^{(-1)}(x) \right| \leq \frac{\varepsilon}{3} \ . \tag{A.8}$$

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

$$\left| \int_{a-\eta_1}^{b+\eta_1} \widetilde{\varphi}_n(x)\, d\Delta_n^{(-1)}(x) - \int_a^b \widetilde{\varphi}(x)\, d\Delta_n^{(-1)}(x) \right| \leq \varepsilon \ .$$

Since ε can be chosen arbitrarily small,

$$\int_{a-\eta_1}^{b+\eta_1} \widetilde{\varphi}_n(x)\, d\Delta_n^{(-1)}(x) - \int_a^b \widetilde{\varphi}(x)\, d\Delta_n^{(-1)}(x) \longrightarrow 0 \ .$$

By using equation (A.2) we get

$$\int_{a-\eta_1}^{b+\eta_1} \widetilde{\varphi}_n(x)\, d\Delta_n^{(-1)}(x) \longrightarrow \int_a^b \widetilde{\varphi}(x)\, d\Delta^{(-1)}(x) \ .$$

Theorem 1.1 of Bai and Silverstein (1998) shows that on 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,

$$\int \widetilde{\varphi}_n(x)\, d\Delta_n^{(-1)}(x) \longrightarrow \int_a^b \widetilde{\varphi}(x)\, d\Delta^{(-1)}(x) \ .$$

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

Lemma A.3.

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

Proof of Lemma A.3.

$$\begin{aligned} \frac{1}{p} \log \Big[\det \big(\Sigma_n^{-1} \widetilde{S}_n \big) \Big] &= \frac{1}{p} \log \Big[\det \big(\Sigma_n^{-1} \big) \det \big(\widetilde{S}_n \big) \Big] \\ &= \frac{1}{p} \log \Big[\det \big(\Sigma_n^{-1} \big) \prod_{i=1}^{p} \widetilde{\varphi}_n(\lambda_i) \Big] \\ &= \int \log \left[\widetilde{\varphi}_n(x) \right] dF_n(x) - \int \log(t)\, dH_n(t) \ . \end{aligned} \tag{A.9}$$

A reasoning analogous to that conducted in the proof of Lemma A.2 shows that the first term on the right-hand side of equation (A.9) converges almost surely to $\int_a^b \log \big[\widetilde{\varphi}(x) \big] dF(x)$. Given that H_n converges weakly to H, Lemma A.3 follows. ■

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

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

Lemma A.3 implies that

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

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

A.2 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 $\eta > 0$, there are no eigenvalues outside the interval $[a-\eta, b+\eta]$ with probability one, for all n large enough.

For any given estimator $\widetilde{S}_n$ with limiting shrinkage function $\widetilde{\varphi}$, define the univariate function $\forall x, y \in [a,b],\ \widetilde{\psi}(x) \coloneqq \widetilde{\varphi}(x)/x$ and 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 A.4.

$$\frac{2}{p^2} \sum_{j=1}^{p} \sum_{i>j} \frac{\lambda_j \widetilde{\psi}(\lambda_j) - \lambda_i \widetilde{\psi}(\lambda_i)}{\lambda_j - \lambda_i} \xrightarrow{\text{a.s.}} \int_a^b \int_a^b \widetilde{\psi}^{\sharp}(x,y)\, dF(x)\, dF(y) \ . \tag{A.10}$$

Proof of Lemma A.4.

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

Given equation (3.1), the first term converges almost surely to the right-hand side of equation (A.10). The absolute value of the second term is bounded by K/p; therefore, it vanishes asymptotically. ■

Lemma A.5.

$$\int_a^b \int_a^b \widetilde{\psi}^{\sharp}(x,y)\, dF(x)\, dF(y) = -2 \int_a^b x\widetilde{\psi}(x)\, \mathsf{Re}\left[\breve{m}_F(x)\right]\, dF(x) \ . \tag{A.11}$$

Proof of Lemma A.5. 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 (A.11) follows. ■

Putting together Lemmas A.4 and A.5 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 A.6. *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 A.6. 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 formulas 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

$$\begin{aligned}&(1-c)\int_a^b \widetilde{\psi}(x)dF(x) - \int_a^b \log[\widetilde{\psi}(x)]dF(x) - 2c\int_a^b x\widetilde{\psi}(x)\mathsf{Re}[\breve{m}_F(x)]dF(x) + \frac{1-c}{c}\log(1-c) \\ &= \int_a^b \left\{\frac{1-c-2\,c\,x\,\mathsf{Re}[\breve{m}_F(x)]}{x}\widetilde{\varphi}(x) - \log[\widetilde{\varphi}(x)]\right\}dF(x) + \int_a^b \log(x)dF(x) + \frac{1-c}{c}\log(1-c) \\ &= \int_a^b \left\{\frac{1-c-2\,c\,x\,\mathsf{Re}[\breve{m}_F(x)]}{x}\widetilde{\varphi}(x) - \log[\widetilde{\varphi}(x)]\right\}dF(x) + \int \log(t)\,dH(t) - 1 \ ,\end{aligned}$$

where the last equality comes from the following lemma.

Lemma A.7. $\displaystyle\int_a^b \log(x)\,dF(x) + \frac{1-c}{c}\log(1-c) = \int \log(t)\,dH(t) - 1$.

Proof of Lemma A.7. Setting $\widetilde{\varphi}(x) = x$ for all $x \in \mathsf{Supp}(F)$ in Lemma A.3 yields

$$\frac{1}{p}\log\Big[\det\left(\Sigma_n^{-1}S_n\right)\Big] \xrightarrow{\text{a.s.}} \int_a^b \log(x)\,dF(x) - \int \log(t)\,dH(t)\ . \tag{A.12}$$

In addition, note that

$$\begin{aligned}\frac{1}{p}\log\Big[\det\left(\Sigma_n^{-1}S_n\right)\Big] &= \frac{1}{p}\log\left[\det\left(\Sigma_n^{-1}\frac{1}{n}\sqrt{\Sigma_n}X_n'X_N\sqrt{\Sigma_n}\right)\right] \\ &= \frac{1}{p}\log\left[\det\left(\frac{1}{n}X_n'X_n\right)\right] \xrightarrow{\text{a.s.}} \frac{c-1}{c}\log(1-c) - 1\ ,\end{aligned} \tag{A.13}$$

where the convergence comes from equation (1.1) of Bai and Silverstein (2004). Comparing equation (A.12) with equation (A.13) proves the lemma. ■

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

A.3 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

$$1 - c - 2\,c\,x_0\,\mathsf{Re}[\breve{m}_F(x_0)] \leq \frac{a_1}{\overline{h}} - 2\varepsilon\ . \tag{A.14}$$

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 $x \in \mathbb{R}$,

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

By Lemmas A.1–A.2,

$$\frac{1}{p}\mathsf{Tr}\left(\Sigma_n^{-1}\overline{S}_n\right) \xrightarrow{\text{a.s.}} \int \overline{\varphi}(x)\,\frac{1 - c - 2\,c\,x\,\mathsf{Re}[\breve{m}_F(x)]}{x}\,dF(x)\ . \tag{A.15}$$

The left-hand side of equation (A.15) is asymptotically bounded from below as follows.

$$\begin{aligned}\frac{1}{p}\mathsf{Tr}\left(\Sigma_n^{-1}\overline{S}_n\right) &= \frac{1}{p}\sum_{i=1}^{p} u_{n,i}'\Sigma_n^{-1}u_{n,i} \times \lambda_{n,i}\,\mathbb{1}_{[x_1,x_2]}(\lambda_{n,i}) \\ &\geq \frac{\lambda_{n,1}}{\overline{h}}\left[F_n(x_2) - F_n(x_1)\right] \xrightarrow{\text{a.s.}} \frac{a_1}{\overline{h}}\left[F(x_2) - F(x_1)\right]\ .\end{aligned} \tag{A.16}$$

The right-hand side of equation (A.15) is bounded from above as follows.

$$\int \overline{\varphi}(x) \, \frac{1 - c - 2\,c\,x\,\mathsf{Re}[\breve{m}_F(x)]}{x} \, dF(x) \leq \left(\frac{a_1}{\overline{h}} - \varepsilon \right) [F(x_2) - F(x_1)] \ . \tag{A.17}$$

Given that $F(x_2) - F(x_1) > 0$, equations (A.15)–(A.17) form a logical contradiction. Therefore, the initial assumption (A.14) must be false, which proves Proposition 3.2. ■

B Proofs of Theorems in Section 4

B.1 Proof of Theorem 4.1

Lemma B.1. *Under Assumptions 3.1–3.3, there exists a nonrandom function $\Delta^{(1)}$ defined on $\mathbb{R}$ such that the random function $\Delta_n^{(1)}(x)$ converges almost surely to $\Delta^{(1)}(x)$, for all $x \in \mathbb{R}$. Furthermore, $\Delta^{(1)}$ is continuously differentiable on $\mathbb{R}$ and can be expressed as*

$$\forall x \in \mathbb{R} \qquad \Delta^{(1)}(x) = \begin{cases} 0 & \text{if } x < a, \\ \int_a^x \delta^{(1)}(\lambda) dF(\lambda) & \text{if } x \geq a, \end{cases}$$

where $\forall \lambda \in [a, +\infty)$, $\delta^{(1)}(\lambda) \coloneqq \lambda / |1 - c - c\,\lambda\,\breve{m}_F(\lambda)|^2$.

Proof of Lemma B.1. Follows directly from Theorem 4 of Ledoit and Péché (2011). ■

Lemma B.2. *Under Assumptions 3.1–3.4,*

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

Proof of Lemma B.2. Note that

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

The remainder of the proof is similar to the proof of Lemma A.2 and is thus omitted. ■

Lemma B.1 and Lemma B.2 imply that

$$\frac{1}{p} \mathsf{Tr}\big(\Sigma_n \widetilde{S}_n^{-1}\big) \xrightarrow{\text{a.s.}} \int_a^b \frac{x}{\widetilde{\varphi}(x) \ |1 - c - c\,x\,\breve{m}_F(x)|^2} \, dF(x) \ . \tag{B.1}$$

Lemma A.3 implies that

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

Putting these two results together completes the proof of Theorem 4.1. ■

B.2 Proof of Theorem 4.2

Note that

$$\begin{aligned} \frac{1}{p} \mathsf{Tr} \left[\Big(\Sigma_n - \widetilde{S}_n \Big)^2 \right] &= \frac{1}{p} \sum_{i=1}^{p} \big[\tau_{n,i}^2 - 2 u_{n,i}' \Sigma_n u_{n,i} \, \widetilde{\varphi}_n(\lambda_{n,i}) + \widetilde{\varphi}_n(\lambda_{n,i})^2 \big] \\ &= \int x^2 \, dH_n(x) - 2 \int \widetilde{\varphi}_n(x) \, d\Delta_n^{(1)}(x) + \int \widetilde{\varphi}_n(x)^2 \, dF_n(x) \ . \end{aligned}$$

The remainder of the proof is similar to the proof of Lemma A.2 and is thus omitted. ■

B.3 Proof of Theorem 4.3

Note that

$$\frac{1}{p}\mathsf{Tr}\left[\left(\Sigma_n^{-1}-\widetilde{S}_n^{-1}\right)^2\right]=\frac{1}{p}\sum_{i=1}^{p}\left[\frac{1}{\tau_{n,i}^2}-2\frac{u_{n,i}'\Sigma_n^{-1}u_{n,i}}{\widetilde{\varphi}_n(\lambda_{n,i})}+\frac{1}{\widetilde{\varphi}_n(\lambda_{n,i})^2}\right]$$
$$=\int\frac{1}{x^2}\,dH_n(x)-2\int\frac{1}{\widetilde{\varphi}_n(x)}\,d\Delta_n^{(-1)}(x)+\int\frac{1}{\widetilde{\varphi}_n(x)^2}\,dF_n(x)\ .$$

The remainder of the proof is similar to the proof of Lemma A.2 and is thus omitted. ■

C Proof of Theorems in Section 5

C.1 Proof of Theorem 5.2

Define the shrinkage function

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

Theorem 2.2 of Ledoit and Wolf (2013) and Proposition 4.3 of Ledoit and Wolf (2012) imply that $\forall x\in\mathsf{Supp}(F)$, $\widehat{\varphi}_n^*(x)\xrightarrow{\text{a.s.}}\varphi^*(x)$, and that this convergence is uniform over $x\in\mathsf{Supp}(F)$, apart from arbitrarily small boundary regions of the support. Theorem 5.2 then follows from Corollary 3.1. ■

D Proof of Theorems in Section 6

D.1 Proof of Theorem 6.1

Lemma D.1. *Under Assumptions 3.2–3.3 and 6.1, there exists a nonrandom function $\Delta^{(-1)}$ defined on $\mathbb{R}$ such that $\Delta_n^{(-1)}(x)$ converges almost surely to $\Delta^{(-1)}(x)$, for all $x\in\mathbb{R}-\{0\}$. Furthermore, $\Delta^{(-1)}$ is continuously differentiable on $\mathbb{R}-\{0\}$ and can be expressed as $\forall x\in\mathbb{R}$, $\Delta^{(-1)}(x)=\int_{-\infty}^{x}\delta^{(-1)}(\lambda)dF(\lambda)$, where*

$$\forall\lambda\in\mathbb{R}\qquad\delta^{(-1)}(\lambda)\coloneqq\begin{cases}0 & \text{if } \lambda<0,\\ \dfrac{c}{c-1}\cdot\breve{m}_H(0)-\breve{m}_{\underline{F}}(0) & \text{if } \lambda=0,\\ \dfrac{1-c-2\,c\,\lambda\,\mathsf{Re}[\breve{m}_F(\lambda)]}{\lambda} & \text{if } \lambda>0\ .\end{cases}$$

Proof of Lemma D.1. The proof of Lemma D.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 6.1. ■

The proof of Theorem 6.1 proceeds as the proof of Theorem 3.1, except that Lemma D.1 replaces Lemma A.1. ■

D.2 Proof of Theorem 6.2

Define the shrinkage function

$$\widehat{\varphi}_n^*(0) \coloneqq \left(\frac{p/n}{p/n - 1} \cdot \widehat{\breve{m}_H}(0) - \widehat{\breve{m}_{\underline{F}}}(0) \right)^{-1},$$

$$\text{and} \quad \forall x \in \mathsf{Supp}\big(\underline{F}^{\widehat{\boldsymbol{\tau}}_n}_{n,p}\big) \qquad \widehat{\varphi}_n^*(x) \coloneqq \frac{x}{1 - \frac{p}{n} - 2\,\frac{p}{n}\,x\,\mathsf{Re}\big[\breve{m}^{\widehat{\boldsymbol{\tau}}_n}_{n,p}(x)\big]}\,.$$

First, since both $\widehat{\breve{m}_H}(0)$ and $\widehat{\breve{m}_{\underline{F}}}(0)$ are strongly consistent estimators, $\widehat{\varphi}_n^*(0) \xrightarrow{\text{a.s}} \varphi^*(0)$. Second, Theorem 2.2 of Ledoit and Wolf (2013) and Proposition 4.3 of Ledoit and Wolf (2012) applied to $\underline{F}$ imply that $\forall x \in \mathsf{Supp}(\underline{F})$, $\widehat{\varphi}_n^*(x) \xrightarrow{\text{a.s}} \varphi^*(x)$, and that this convergence is uniform over $x \in \mathsf{Supp}(\underline{F})$, apart from arbitrarily small boundary regions of the support. Theorem 6.2 then follows from Corollary 6.1. ■