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

Ledoit, Olivier ; Wolf, Michael

Abstract: This paper introduces a new method for deriving covariance matrix estimators that are decision-theoretically optimal within a class of nonlinear shrinkage estimators. 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 (Estimation of a covariance matrix (1975); J. Math. Sci. 34 (1986) 1373–1403), 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 via 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's 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 in the large-dimensional asymptotic limit with respect to the Frobenius loss function.

DOI: https://doi.org/10.3150/17-bej979

Bernoulli **24**(4B), 2018, 3791–3832
https://doi.org/10.3150/17-BEJ979

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

OLIVIER LEDOIT[1,2,*,**] and MICHAEL WOLF[1,†]

[1]*Department of Economics, University of Zurich, 8032 Zurich, Switzerland.*
E-mail: [*] *olivier.ledoit@econ.uzh.ch;* [†] *michael.wolf@econ.uzh.ch*
[2]*AlphaCrest Capital Management, New York, NY 10036, USA.*
E-mail: [**] *olivier.ledoit@alphacrestcapital.com*

This paper introduces a new method for deriving covariance matrix estimators that are decision-theoretically optimal within a class of nonlinear shrinkage estimators. 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 (Estimation of a covariance matrix (1975); *J. Math. Sci.* **34** (1986) 1373–1403), 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 via 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's 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 in the large-dimensional asymptotic limit with respect to the Frobenius loss function.

Keywords: large-dimensional asymptotics; nonlinear shrinkage estimation; random matrix theory; rotation equivariance; Stein's loss

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 [41,42]. 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 "shrink-

1350-7265

ing") 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 [16], Lin and Perlman [30], Dey and Srinivasan [12], Daniels and Kass [10], Ledoit and Wolf [26,27], Chen, Wiesel and Hero [9], Won et al. [44], and the references therein.

Although Stein's estimator is still considered the "gold standard" (Rajaratnam and Vincenzi [36]) 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 [41,42] 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 [30], 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 [41, 42] 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 [42] 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 our large-dimensional framework, and in the class of nonlinear shrinkage estimators under consideration. 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. In particular, concerning point 4., Stein [41,42] assumes normality to show that the relevant objective function is an unbiased estimator of the risk. But as we establish in the present paper, this objective function converges to the same limit as the risk in an appropriate asymptotic setting. Hence, our results demonstrate that Stein's estimator – as well as our new estimator – is also a relevant estimator when normality does not hold.

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 the oracle's asymptotic optimality property. The second step is not original, as we adapt technology developed earlier by Ledoit and Wolf [27,28]. 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 [27,28] 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 [27,28].

Additional evidence for our method being effective is the fact that it enables us to discover a new oracle covariance matrix estimator that is optimal with respect to the "Symmetrized Stein's loss" within our class of nonlinear shrinkage estimators, under large-dimensional asymptotics. 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 [41,42]. 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 in the nonlinear shrinkage class with respect to various loss functions. Section 5 develops our feasible estimator of a covariance matrix, which is asymptotically optimal in the nonlinear shrinkage class 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 investigates the case where the largest eigenvalue goes to infinity at the same rate as the matrix dimension while the bulk of the eigenvalues remain bounded. Section 8 studies finite-sample properties via Monte Carlo simulations. Section 9 shows an empirical application to real data. Section 10 contains concluding remarks. The proofs of all mathematical results are collected in the supplementary material (Ledoit and Wolf [29]).

2. Shrinkage in finite samples under Stein's loss

This section expounds the finite-sample theory of Stein [41,42], 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 $\boldsymbol{\tau}_n := (\tau_{n,1}, \ldots, \tau_{n,p})'$ denote a system of eigenvalues of Σ_n. The empirical distribution function (e.d.f.) of the population eigenvalues is defined as

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

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 $Y_n := X_n \times \sqrt{\Sigma_n}$, where $\sqrt{\ }$ denotes the symmetric positive-definite square root of a 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 $S_n := n^{-1} Y_n' Y_n = n^{-1} \sqrt{\Sigma_n} X_n' X_n \sqrt{\Sigma_n}$. 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 $\Lambda_n := \mathsf{Diag}(\boldsymbol{\lambda}_n)$ where $\boldsymbol{\lambda}_n := (\lambda_{n,1}, \ldots, \lambda_{n,p})'$. 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 $\widetilde{S}_n := U_n \widetilde{D}_n U_n'$, where $\widetilde{D}_n$ is a diagonal matrix: $\widetilde{D}_n := \mathsf{Diag}(\widetilde{\varphi}_n(\lambda_{n,1}) \ldots, \widetilde{\varphi}_n(\lambda_{n,p}))$, and $\widetilde{\varphi}_n$ is a (possibly random) real univariate function which can depend on S_n.

(Since $\widetilde{\varphi}_n$ is allowed to depend on S_n, in particular, $\widetilde{\varphi}_n(\lambda_{n,i})$ is not necessarily a function of $\lambda_{n,i}$ only but may depend on the other $\lambda_{n,j}$ also.)

This is the class of "rotation-equivariant" estimators introduced by Stein [41,42]: 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 [42] does not work with the function $\widetilde{\varphi}_n(\cdot)$ itself but with the vector $(\widetilde{\varphi}_{n,1}, \ldots, \widetilde{\varphi}_{n,p})' := (\widetilde{\varphi}_n(\lambda_{n,1}), \ldots, \widetilde{\varphi}_n(\lambda_{n,p}))'$ instead. This is equivalent because the sample eigenvalues are distinct with probability one, and because the values taken by the shrinkage function $\widetilde{\varphi}_n(\cdot)$ 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 RMT literature have found it more tractable to work with the e.d.f. of the sample eigenvalues,

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

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 [41,42] and commonly referred to as Stein's loss:

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

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

Note that Stein [41,42] 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".

Stein's loss is proportional to the Kullback and Leibler [24] divergence from the multivariate normal distribution with zero mean and covariance matrix Σ_n to the multivariate normal distribution with zero mean and covariance matrix $\widetilde{S}_n$, which is commonly expressed in the following notation:

$$\mathcal{L}_n^S(\Sigma_n, \widetilde{S}_n) = \frac{2}{p} D_{\mathrm{KL}}\big(\mathcal{N}(0, \widetilde{S}_n) \parallel \mathcal{N}(0, \Sigma_n)\big). \tag{2.1}$$

2.2. Stein's loss in finite samples

Stein [42] introduces a function closely related to the nonlinear shrinkage function: $\widetilde{\psi}(x) := \widetilde{\varphi}(x)/x$. Under Assumptions 2.1–2.3, Stein shows that the risk function satisfies the identity $\mathcal{R}_n^S(\Sigma_n, \widetilde{S}_n) = \mathbb{E}[\Theta_n(\Sigma_n, \widetilde{S}_n)]$, with

$$\begin{aligned}\Theta_n(\Sigma_n, \widetilde{S}_n) := {} & \frac{n-p+1}{np} \sum_{j=1}^{p} \widetilde{\psi}_n(\lambda_{n,j}) - \frac{1}{p} \sum_{j=1}^{p} \log\big[\widetilde{\psi}_n(\lambda_{n,j})\big] + \log(n) \\ & + \frac{2}{np} \sum_{j=1}^{p} \sum_{i>j} \frac{\lambda_{n,j}\widetilde{\psi}_n(\lambda_{n,j}) - \lambda_{n,i}\widetilde{\psi}_n(\lambda_{n,i})}{\lambda_{n,j} - \lambda_{n,i}} \\ & + \frac{2}{np} \sum_{j=1}^{p} \lambda_{n,j} \widetilde{\psi}_n'(\lambda_{n,j}) - \frac{1}{p} \sum_{j=1}^{p} \mathbb{E}\big[\log\big(\chi^2_{n-j+1}\big)\big] - 1,\end{aligned} \tag{2.2}$$

where

$$\widetilde{\psi}_n'(x) := \frac{\partial \widetilde{\psi}_n(x)}{\partial x}$$

and χ^2_{n-j+1}, for $j = 1, \ldots, p$, denote independent chi-square random variables with respective degrees of freedom as indicated by their subscripts; for example, see Muirhead [34], Theorems 3.2.15 and 4.3.1. 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

$$\forall i = 1, \ldots, p \qquad \varphi_n^*(\lambda_{n,i}) := \frac{\lambda_{n,i}}{1 - \frac{p-1}{n} - 2\frac{p}{n}\lambda_{n,i} \times \frac{1}{p}\sum_{j \neq i} \frac{1}{\lambda_{n,j} - \lambda_{n,i}}}. \tag{2.3}$$

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 (2.3)

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 and Vincenzi [36]. Eschewing isotonization is one of our motivations for going to large-dimensional asymptotics.

The appendix of Lin and Perlman [30] gives a detailed description of the isotonizing algorithm. If we call the isotonized shrinkage function φ_n^{ST}, Stein's "isotonized" estimator is

$$S_n^{\mathrm{ST}} := U_n D_n^{\mathrm{ST}} U_n', \qquad \text{where } D_n^{\mathrm{ST}} := \mathsf{Diag}\big(\varphi_n^{\mathrm{ST}}(\lambda_{n,1}), \ldots, \varphi_n^{\mathrm{ST}}(\lambda_{n,p})\big). \tag{2.4}$$

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.

a. The population covariance matrix Σ_n is a nonrandom symmetric positive-definite matrix of dimension $p \times p$.
b. 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)".
c. $\mathsf{Supp}(H)$, the support of H, is the union of a finite number of closed intervals, bounded away from zero and infinity.
d. There exists a compact interval $[\underline{h}, \overline{h}] \subset (0, \infty)$ that contains $\{\tau_{n,1}, \ldots, \tau_{n,p}\}$ 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 [1] for a comprehensive review. The assumption that $\mathsf{Supp}(H_n)$ is uniformly bounded away from zero is widespread and made by such authors as Johnstone [22], Bickel and Levina [8], Mestre [32], Won et al. [44], and Khare, Oh and Rajaratnam [23], among others. The assumption that $\mathsf{Supp}(H_n)$ is uniformly bounded away from infinity is even more widespread and made by

such authors as Bai and Silverstein [2–4], Johnstone [22], Bickel and Levina [8], Mestre [32], El Karoui [14], Won et al. [44], and Khare, Oh and Rajaratnam [23], among others. In particular, our set of assumptions is much less restrictive than the "spike model" of Johnstone [22] which is still widely in use; for example, see Donoho, Gavish and Johnstone [13]. (Note that Bickel and Levina [8] use only the assumption of an upper bound for estimating the covariance matrix itself, whereas they use the assumption of both a lower and an upper bound for estimating the inverse of the covariance matrix.)

Furthermore, since in Assumption 3.2.d the lower bound $\underline{h}$ can be arbitrarily small and the upper bound $\overline{h}$ can be arbitrarily large, the assumption also covers the case of a poorly conditioned covariance matrix. Indeed, Monte Carlo simulations reported in Figure 3 indicate that our estimator performs well in practice even when the smallest eigenvalue goes to zero, while Monte Carlo simulations reported in Figure 9 indicate that our estimator performs well in practice even when the largest eigenvalue goes to infinity.

In order to streamline the language, we adopt the convention throughout the paper that the words "limit", "convergence", "asymptotic", and variations thereof, signify convergence under large-dimensional asymptotics as defined by Assumptions 3.1–3.2, unless explicitly stated otherwise.

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 $Y_n := X_n \times \sqrt{\Sigma_n}$. 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 [25]. However, Monte Carlo studies in Ledoit and Wolf [27,28] 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 [31]. It has been strengthened and broadened by subsequent authors including Silverstein [37], Silverstein and Bai [38], and Silverstein and Choi [39], among others. These works 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 [37] is that $m := m_F(z)$ is the unique solution in the set

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

to the equation

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

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:

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

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 [39] 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 [2], 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 $0 < a_1 < b_1 < \cdots < a_\kappa < b_\kappa < \infty$.

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.

Remark 3.2. The uniform convergence in Assumption 3.4 means that for any small $\eta > 0$, there exists a set of probability one such that on this set, $\sup_{x \in \mathcal{A}_\eta} |\widetilde{\varphi}_n(x) - \widetilde{\varphi}(x)| \longrightarrow 0$, with $\mathcal{A}_\eta := \bigcup_{k=1}^{\kappa} [a_k + \eta, b_k - \eta]$. This assumption is used in the proof of Lemma 11.2 in the supplementary material Ledoit and Wolf [29].

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 $\widetilde{\varphi}$.

Throughout the paper, we call the set of estimators specified by Definition 2.1 and Assumption 3.4 "the class of nonlinear shrinkage estimators". We argue that this is not a restrictive definition for two reasons: first, for finite dimension p and sample size n, the shrunk eigenvalues $(\widetilde{\varphi}_n(\lambda_{n,1}) \ldots, \widetilde{\varphi}_n(\lambda_{n,p}))$ can be anything in $\mathbb{R}^n$; second, all we require is that the shrinkage function $\widetilde{\varphi}_n$ remains bounded and converges uniformly to some continuously differentiable limit $\widetilde{\varphi}$. It would be very difficult to deal mathematically with shrinkage functions that are unbounded, or that alternate between vastly different shapes without ever converging to any specific one.

3.2. Stein's loss under large-dimensional asymptotics

Instead of minimizing the unbiased estimator of risk $\Theta_n(\Sigma_n, \widetilde{S}_n)$ defined in Equation (2.2), as Stein [41,42] does, we minimize $\lim_{p,n\to_c\infty} \mathcal{L}_n^S(\Sigma_n, \widetilde{S}_n)$, where the loss function $\mathcal{L}_n^S$ comes from Definition 2.2, 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 positive constant c; see Assumption 3.1.

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

Theorem 3.1. *Under Assumptions 3.1–3.4,*

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

The proof is in Section 11.1 of the supplementary material Ledoit and Wolf [29].

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,*

$$\Theta_n(\Sigma_n, \widetilde{S}_n) \xrightarrow{\text{a.s.}} \mathcal{M}_c^S(H, \widetilde{\varphi}). \tag{3.6}$$

The proof is in Section 11.2 of the supplementary material Ledoit and Wolf [29]. 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.*

a. *A covariance matrix estimator $\widetilde{S}_n$ minimizes in the class of rotation-equivariant estimators described in Definition 2.1 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*

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

The resulting oracle estimator of the covariance matrix is

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

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

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

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 Section 11.3 of the supplementary material Ledoit and Wolf [29].

Proposition 3.2. *Under Assumptions* 3.1–3.3,

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

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.

3.3. Comparison with other estimators

The techniques developed above are sufficiently general to enable us to compute the almost sure limit of Stein's loss for other covariance matrix estimators as well. Countless estimators of the covariance matrix have been proposed in the literature and it is well beyond the scope of the present paper to review them all. We restrict attention to three estimators that will be included in the Monte Carlo simulations of Section 8 and the in empirical application of Section 9.

3.3.1. *Sample covariance matrix*

The sample covariance matrix fits in our framework by taking the shrinkage function: $\varphi_n^S(x) := x$, for all $x \in \mathbb{R}$. It converges to the limiting shrinkage function $\varphi^S(x) := x$ uniformly over $\mathbb{R}$.

Applying Theorem 3.1 yields the almost sure limit of Stein's loss for the sample covariance matrix:

$$\begin{aligned}\mathcal{L}_n^S(\Sigma_n, S_n) &\xrightarrow{\text{a.s.}} \mathcal{M}_c^S\big(H, \varphi^S\big) \\ &= \int \log(t)\, dH(t) - \sum_{k=1}^{\kappa} \int_{a_k}^{b_k} \big\{c + 2cx\, \mathsf{Re}\big[\breve{m}_F(x)\big] + \log(x)\big\}\, dF(x).\end{aligned} \tag{3.9}$$

Corollary 3.1 implies that the limiting loss of the sample covariance matrix $\mathcal{M}_c^S(H, \varphi^S)$ is at least as high as that of the optimal nonlinear shrinkage function $\mathcal{M}_c^S(H, \varphi^*)$. However, it may be possible *a priori* that the losses are equal for some parameter configurations. By directly comparing Equations (3.8) and (3.9), we can establish that this is nowhere the case.

Proposition 3.3. *For any $c \in (0, 1)$ and any cumulative distribution function H satisfying Assumption 3.2.c, $\mathcal{M}_c^S(H, \varphi^S) > \mathcal{M}_c^S(H, \varphi^*)$.*

3.3.2. *Minimax estimator*

Theorem 3.1 of Dey and Srinivasan [12] presents a covariance matrix estimator that is minimax with respect to Stein's loss within the class of rotation-equivariant estimators specified in Definition 2.1. These authors acknowledge that the same estimator was presented by Charles Stein in a series of lectures given at the University of Washington, Seattle in 1982. Their minimax estimator is obtained by multiplying the ith sample eigenvalue by the coefficient

$$\Delta_i := \frac{n}{n + p + 1 - 2(p + 1 - i)}, \qquad i = 1, \ldots, p. \tag{3.10}$$

In terms of notation, the term $(p + 1 - i)$ in the denominator appears in the original paper as i because Dey and Srinivasan [12] sort eigenvalues in descending order, whereas we use the convention that they are sorted in ascending order. Also, we need to introduce the quantity n in the denominator because Dey and Srinivasan [12] work with the eigenvalues of $n \times S_n$, whereas we work with the eigenvalues of S_n.

The coefficient Δ_i from Equation (3.10) has a long history, having been originally introduced by Stein [40], Equation (4.11), and James and Stein [20], Equation (85), in the context of minimax estimators of the covariance matrix that are *not* rotation-equivariant. We can rewrite Δ_i as

$$\Delta_i = \frac{n}{n - p - 1 + 2pF_n(\lambda_{n,i})}, \qquad i = 1, \ldots, p. \tag{3.11}$$

Therefore, the minimax shrinkage function is defined in finite samples as

$$\forall x \in \mathbb{R} \qquad \varphi_n^M(x) := \frac{x}{1 - \frac{p+1}{n} + 2\frac{p}{n}F_n(x)}, \tag{3.12}$$

and converges almost surely to the limiting shrinkage function

$$\forall x \in \mathbb{R} \qquad \varphi^M(x) := \frac{x}{1 - c + 2cF(x)} \tag{3.13}$$

uniformly over the support of F. Plugging φ^M into Theorem 3.1 yields the almost sure limit of Stein's loss for the minimax estimator:

$$\begin{aligned}\mathcal{M}_c^S(H,\varphi^M) := &\int \left\{ \frac{1-c-2cx\,\mathsf{Re}[\breve{m}_F(x)]}{1-c+2cF(x)} - \log\left[\frac{x}{1-c+2cF(x)}\right]\right\} dF(x) \\ &+ \int \log(t)\,dH(t) - 1.\end{aligned} \tag{3.14}$$

As in Section 3.3.1 above, Corollary 3.1 implies that the limiting loss of the minimax estimator $\mathcal{M}_c^S(H,\varphi^M)$ is at least as high as that of the optimal nonlinear shrinkage function $\mathcal{M}_c^S(H,\varphi^*)$. However, it may be possible *a priori* that the losses are equal for some parameter configurations. By directly comparing Equations (3.8) and (3.14), we can establish that this is nowhere the case.

Proposition 3.4. *For any $c \in (0,1)$ and any cumulative distribution function H satisfying Assumption 3.2.c, $\mathcal{M}_c^S(H,\varphi^M) > \mathcal{M}_c^S(H,\varphi^*)$.*

Thus, our estimator strictly improves pointwise upon the minimax estimator according to Stein's loss, which implies that the minimax estimator is inadmissible in the large-dimensional asymptotic limit.

3.3.3. *Linear shrinkage*

Let φ_n^L denote the linear shrinkage formula of Ledoit and Wolf [26], Equation (14).

Proposition 3.5. *The linear shrinkage function φ_n^L converges almost surely to*

$$\varphi^L : x \longmapsto \int \lambda\,dF(\lambda) + \frac{\int t^2\,dH(t) - [\int t\,dH(t)]^2}{\int \lambda^2\,dF(\lambda) - [\int \lambda\,dF(\lambda)]^2}\left[x - \int \lambda\,dF(\lambda)\right] \tag{3.15}$$

uniformly over the support of F.

Any interested reader can obtain the almost sure limit of Stein's loss for the optimal linear shrinkage estimator $\mathcal{M}_c^S(H,\varphi^L)$ simply by plugging Equation (3.15) into Theorem 3.1. The resulting formula is cumbersome, so we omit it to save space. By Corollary 3.1, $\mathcal{M}_c^S(H,\varphi^L)$ is always at least as high as the limiting loss of the optimal nonlinear shrinkage function $\mathcal{M}_c^S(H,\varphi^*)$. There are some special cases where the two limiting losses may be equal: These are the cases where the optimal nonlinear shrinkage function 'happens' to be exactly linear; one such case is when all population eigenvalues are equal to one another. However, in the generic case, nonlinear shrinkage is strictly better asymptotically, since linear shrinkage estimators form a two-dimensional subspace nested inside the p-dimensional space of nonlinear shrinkage estimators (where p is arbitrarily large).

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 [27,28] 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 [27,28].

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 [43] thus defines the loss function

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

Its limit is given by the following theorem, whose proof is in Section 12.1 of the supplementary material Ledoit and Wolf [29].

Theorem 4.1. *Under Assumptions 3.1–3.4,*

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

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 2.1 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*

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

4.2. Frobenius loss

Ledoit and Wolf [27,28] use the following loss function based on the squared Frobenius distance:

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

Its limit is given by the following theorem, whose proof is in Section 12.2 of the supplementary material Ledoit and Wolf [29].

Theorem 4.2. *Under Assumptions 3.1–3.4,*

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

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 2.1 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 [27], Section 3.1, and Ledoit and Wolf [28], 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 [27], Section 3.2, apply the Frobenius loss to the precision matrix:

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

Its limit is given by the following theorem, whose proof is in Section 12.3 of the supplementary material Ledoit and Wolf [29].

Theorem 4.3. *Under Assumptions 3.1–3.4,*

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

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 2.1 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 [33].[1] The "Symmetrized Stein's loss (function)" is defined as

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

This loss function is symmetric in the sense that $\mathcal{L}_n^{\mathrm{SSYM}}(\Sigma_n, \widetilde{S}_n) = \mathcal{L}_n^{\mathrm{SSYM}}(\Sigma_n^{-1}, \widetilde{S}_n^{-1})$, and also in the sense that $\mathcal{L}_n^{\mathrm{SSYM}}(\Sigma_n, \widetilde{S}_n) = \mathcal{L}_n^{\mathrm{SSYM}}(\widetilde{S}_n, \Sigma_n)$. It is equal to the Jeffreys [21] divergence between the multivariate normal distribution with zero mean and covariance matrix Σ_n and the multivariate normal distribution with zero mean and covariance matrix $\widetilde{S}_n$, rescaled by the factor $1/p$. Its limit is given by the following theorem.

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

Theorem 4.4. *Under Assumptions 3.1–3.4,*

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

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 2.1 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 $S_n^{\circledast} := U_n \times \mathsf{Diag}(\varphi^{\circledast}(\lambda_{n,1}), \ldots, \varphi^{\circledast}(\lambda_{n,p})) \times U_n'$. 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 $\forall x \in \mathbb{R}$, $\varphi^*(x) < \varphi^{\circledast}(x) < \varphi^\circ(x)$.

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 [27,28] 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 [33] 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 is the square of a distance; see Moakher and Batchelor [33], page 288.

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 [27,28]. 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 [28] 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

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

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

where

$$\forall i = 1,\ldots,p \qquad q^i_{n,p}(\mathbf{t}) := p\int_{(i-1)/p}^{i/p} \bigl(F^{\mathbf{t}}_{n,p}\bigr)^{-1}(u)\,du, \tag{5.3}$$

$$\forall u \in [0,1] \qquad \bigl(F^{\mathbf{t}}_{n,p}\bigr)^{-1}(u) := \sup\bigl\{x \in \mathbb{R} : F^{\mathbf{t}}_{n,p}(x) \le u\bigr\}, \tag{5.4}$$

$$\forall x \in \mathbb{R} \qquad F^{\mathbf{t}}_{n,p}(x) := \lim_{\eta\to 0^+} \frac{1}{\pi}\int_{-\infty}^{x} \mathsf{Im}\bigl[m^{\mathbf{t}}_{n,p}(\xi+i\eta)\bigr]\,d\xi, \tag{5.5}$$

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

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

to the equation

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

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, $F^{\mathbf{t}}_{n,p}$ is the limiting distribution (function) of sample eigenvalues corresponding to the population spectral distribution (function)

$p^{-1}\sum_{i=1}^{p}\mathbb{1}_{[t_i,+\infty)}$. Furthermore, by Equation (5.4), $(F_{n,p}^{\mathbf{t}})^{-1}$ 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 [28] 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*

$$\widehat{\boldsymbol{\tau}}_n := \underset{\mathbf{t}\in(0,\infty)^p}{\arg\min}\,\frac{1}{p}\sum_{i=1}^{p}\bigl[q_{n,p}^{i}(\mathbf{t})-\lambda_{n,i}\bigr]^2, \tag{5.8}$$

where $\boldsymbol{\lambda}_n := (\lambda_{n,1},\ldots,\lambda_{n,p})'$ *are the eigenvalues of the sample covariance matrix* S_n, *and* $Q_{n,p}(\mathbf{t}) := (q_{n,p}^{1}(\mathbf{t}),\ldots,q_{n,p}^{p}(\mathbf{t}))'$ *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* $\boldsymbol{\tau}_n := (\tau_{n,1},\ldots,\tau_{n,p})'$ *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 [28], Theorem 2.2. The solution to Equation (5.8) can be found by standard nonlinear optimization software such as SNOPT™ (Gill, Murray and Saunders [15]) or the MATLAB™ Optimization Toolbox.

5.3. Asymptotically optimal estimator of the covariance matrix

Recall that, for any $\mathbf{t} := (t_1,\ldots,t_p)' \in (0,+\infty)^p$, 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 [39] 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 Section 13 of the supplementary material Ledoit and Wolf [29].

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

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

minimizes in the class of rotation-equivariant estimators described in Definition 2.1 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 in Ledoit and Wolf [28]. 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 $\widehat{S}^\circledast := \sqrt{\widehat{S}^* \widehat{S}^\circ}$. 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 [41] 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 a "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

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

Henrici [18], pages 259–262, is a useful reference for principal values. Stein's shrinkage function and ours – Equations (2.3) and (5.9), respectively – can be expressed as

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

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

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

Both F_n and $F_{n,p}^{\widehat{\boldsymbol{\tau}}_n}$ 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 8 will confirm.

On a more anecdotal note, the shrinkage function of the minimax estimator in Theorem 3.1 of Dey and Srinivasan [12] can also be expressed in nearly identical format as

$$\forall i = 1, \ldots, p \qquad \varphi_n^M(\lambda_{n,i}) = \frac{\lambda_{n,i}}{1 - \frac{p+1}{n} + 2\frac{p}{n} \times F_n(\lambda_{n,i})}. \tag{5.10}$$

We can see that the overall pattern is surprisingly similar, except for the fact that the empirical sample spectral distribution $F_n(x)$ acts as a substitute for the function $x \mapsto \mathrm{PV} \int \frac{x}{x-\lambda} \, d\widehat{F}_{n,p}^{\widehat{\tau}_n}(x)$. The only evident common points are that both functions take the value zero at $x = 0$, and they both converge to the limit one as x goes to infinity.

5.4. Comparison with other approaches from decision theory

Given that we claim our estimator is "decision-theoretically optimal" in a sense that is not completely standard, it is important to compare and constrast our approach with the rest of the literature on decision-theoretical estimation of the covariance matrix.

5.4.1. *Commonalities*

The first common point is that our approach is firmly rooted in decision theory in the sense that the decision (choice of estimator) depends on the loss function: Stein's Loss, Stein's Inverse Loss, and Stein's Symmetrized Loss all lead to different estimators. This has always been a central feature of decision-theoretic estimation, and we are no exception. Thus, the estimator given in Theorem 5.2 is more properly referred to as "decision-theoretically optimal with respect to Stein's Loss".

The second common point is that, in keeping with a long tradition in decision-theoretic estimation of the covariance matrix, we consider only rotation-equivariant estimators that are obtained by manipulating the sample eigenvalues, while preserving the sample eigenvectors. This manipulation is operated by what we call the shrinkage function $\widetilde{\varphi}_n$ and, for fixed n, is unconstrained.

Up to this point, any reader steeped in decision-theoretic estimation of the covariance matrix is still in familiar territory.

5.4.2. *Key difference*

The key difference is that we do not work in finite samples but in the large-dimensional asymptotic limit, where the ratio of dimension to sample size converges to some limit $c > 0$. This approach has a number of consequences that need to be spelled out.

First, manipulating the covariance matrix itself becomes difficult, since its dimension keeps changing and goes to infinity. This is why – instead of working directly with the initial object of

interest, the covariance matrix – we work with its eigenvalues and, more precisely, with limiting spectral distributions. Doing so requires spectral distributions to have well-defined, nonstochastic limits. In the standard setup of random matrix theory, which we adopt, the spectral distribution of the population covariance matrix converges to a well-defined, nonstochastic limit H; and so does the spectral distribution of the sample covariance matrix. Here the only restriction is that we limit ourselves (for mathematical reasons) to population covariance matrices whose condition number does not explode with the dimension. This point is discussed in depth below Assumption 3.2.

Second, manipulating an *estimator* of the covariance matrix also becomes difficult, for the same reasons as above. This is why we introduce in Assumption 3.4 the notion of a limiting shrinkage function $\widetilde{\varphi}$: It guarantees that the spectral distribution of the *shrunk* covariance matrix also has a well-defined, nonstochastic limit. It would be hard to see how we could proceed otherwise. The only restrictions are that the nonlinear shrinkage function must remain bounded, that the convergence must be uniform, and that the limiting shrinkage function must be continuously differentiable. Making these relatively reasonable technical assumptions is what enables us to derive sweeping results.

5.4.3. *Relation to minimax*

A pervasive concept in decision theory is that of a minimax estimator. This means that the estimator $\widetilde{\varphi}$ minimizes the worst-case risk $\sup_H \mathcal{M}_c^S(H, \widetilde{\varphi})$ in the class of estimators considered. Such an approach is justified because in general the risk cannot be directly minimized, since it depends on the unknown parameter itself (which, in this case, is H).

Our situation here is completely different: H can be estimated consistently and hence, asymptotically, the risk can be directly minimized. Indeed this is precisely what Theorem 5.2 says. Thus, it would be misleading to describe our estimator as minimax: A more accurate characterization is that it is pointwise optimal for any H, or uniformly better for all H. This is, obviously, a stronger notion of decision-theoretic optimality than minimax, and one that is generally unattainable in finite samples.

6. Extension to the singular case

So far, we have only considered the case $p < n$, as does Stein [41,42]. We do not know whether Stein was uninterested in the 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

[28], Section 2.1. Define

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

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 [2], 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,*

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

The proof is in Section 14.1 of the supplementary material Ledoit and Wolf [29]. 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 2.1 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*

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

The resulting oracle estimator of the covariance matrix is

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

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

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

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

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

As explained in Ledoit and Wolf [28], 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*

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

$$\forall i = 1, \ldots, p-n \qquad \widehat{\varphi}_n^*(\lambda_{n,i}) := \left(\frac{p/n}{p/n-1}\cdot \widehat{\breve{m}_H}(0) - \widehat{\breve{m}_{\underline{F}}}(0)\right)^{-1}, \tag{6.6}$$

$$\textit{and} \quad \forall i = p-n+1, \ldots, p \qquad \widehat{\varphi}_n^*(\lambda_{n,i}) := \frac{\lambda_{n,i}}{1-\frac{p}{n}-2\frac{p}{n}\lambda_{n,i}\,\mathsf{Re}\big[\breve{m}_{n,p}^{\widehat{\boldsymbol{\tau}}_n}(\lambda_{n,i})\big]} \tag{6.7}$$

minimizes in the class of rotation-equivariant estimators described in Definition 2.1 *the almost sure limit* (6.1) *of Stein's loss.*

The proof is in Section 14.2 of the supplementary material Ledoit and Wolf [29].

Remark 6.1 (Case $p = n$). We have treated the cases $p < n$ and $p > n$. The remaining case $p = n$ cannot be treated theoretically, since a large number of fundamental results from the RMT literature used in our proofs rule out the case $c = 1$, where c is recalled to be the limiting concentration; see Assumptions 3.1 and 6.1. Nevertheless, we can address the case $p = n$ in Monte Carlo simulations; see Figure 2.

7. The arrow model

In common with a large portion of the existing literature, Assumption 3.1 requires the largest population eigenvalue to remain bounded. There are some applications where this may be unrealistic. In this section, we investigate what happens when the largest eigenvalue goes to infinity at the same rate as the dimension and the sample size, while the bulk of the eigenvalues remain bounded.

7.1. Specification

In keeping with standard nomenclature, we call the eigenvalues that remain bounded the "bulk". To distinguish our model from Johnstone's [22] "spike" model, where the largest eigenvalues remains bounded, we call the eigenvalues that shoot up to infinity "arrows". Therefore, Assumption 3.2.d becomes:

Assumption 3.2.e (Arrow model). There exists a compact interval $[\underline{h}, \overline{h}] \subset (0, \infty)$ that contains the set $\{\tau_{n,1}, \ldots, \tau_{n,p-k}\}$ for all n large enough, where k is a fixed integer. There exist k constants $(\beta_j)_{j=1,\ldots,k}$ with $0 < \beta_1 < \cdots < \beta_k$ s.t. $\forall j = 1, \ldots, k$, $\tau_{n,p-k+j} \sim \beta_j p$.

We consider only values of n and p large enough so that the ordering of the arrow eigenvalues $(\tau_{n,p-k+j})_{j=1,\ldots,k}$ matches the ordering of the slopes $(\beta_j)_{j=1,\ldots,k}$.

This is challenging because the papers by Yin, Bai and Krishnaiah [45], Bai, Silverstein and Yin [5], Johnstone [22], Baik, Ben Arous and Péché [6], and Baik and Silverstein [7] that study the asymptotic behavior of the largest sample eigenvalue all assume it to be bounded. Given the dearth of background results applicable to the arrow model, this section is (by necessity) exploratory in nature. Until the underlying probability theory literature has caught up, the robustness of Theorem 5.2 against Assumption 3.2.e must remain a conjecture.

Nevertheless, we can make some significant inroads by resorting to alternative methods such as the Weyl inequalities and perturbation theory. Given that this investigation plays only a supporting role relative to the main contributions of the paper, and that even the most basic properties have to be established from scratch, we restrict ourselves to the single-arrow case: $k = 1$.

Assumption 3.2.f (Single-arrow model). There exist a compact interval $[\underline{h}, \overline{h}] \subset (0, \infty)$ that contains the set $\{\tau_{n,1}, \ldots, \tau_{n,p-1}\}$ for all n large enough, and a constant $\beta_1 > 0$ s.t. $\tau_{n,p} \sim \beta_1 p$.

This section presents a collection of propositions that, together, indicate that the single-arrow model is no particular cause for concern. The basic intuition is that the arrow sticks out like a

sore thumb in any data set of sufficient size. Therefore, it is easy to detect its presence, separate it from the bulk, measure its variability (eigenvalue), find its orientation (eigenvector), apply an appropriate amount of shrinkage to it, partial it out, and then deal with the bulk as usual. We present preliminary evidence suggesting that our proposed estimator $\widehat{S}_n^*$ does all of the above automatically.

7.2. Spectral separation

All the proofs from this section are in Section 15 of the supplementary material Ledoit and Wolf [29]. Our first proposition shows that the bulk sample eigenvalues $(\lambda_{n,1}, \ldots, \lambda_{n,p-1})$ remain bounded, while the arrow sample eigenvalue $\lambda_{n,p}$ goes to infinity.

Proposition 7.1. *Under Assumptions* 3.1, 3.2.*a–c*, 3.2.*f and* 3.3, $\lambda_{n,p-1}$ *remains bounded a.s. for large* n, *and* $\lambda_{n,p} \xrightarrow{\text{a.s.}} \infty$.

It means that we observe what RMT calls "spectral separation" between the bulk and the arrow. The size of the gap grows arbitrarily large. The good news is that the QuEST function automatically follows the same pattern of spectral separation.

Proposition 7.2. *Under Assumptions* 3.1, 3.2.*a–c and* 3.2.*f*, $q_{n,p}^{p-1}(\boldsymbol{\tau}_n)$ *remains bounded and* $q_{n,p}^{p}(\boldsymbol{\tau}_n) \to \infty$.

The similarity between Proposition 7.1 and Proposition 7.2 gives reassurance about the ability of the QuEST function (5.2) to separate the arrow from the bulk.

7.3. Sample arrow eigenvalue

Our next proposition shows that the arrow sample eigenvalue is asymptotically equivalent to its population counterpart.

Proposition 7.3. *Under Assumptions* 3.1, 3.2.*a–c*, 3.2.*f and* 3.3, $\lambda_{n,p} \overset{\text{a.s.}}{\sim} \tau_{n,p}$.

It is surprising that the sample arrow eigenvalue is asymptotically equivalent to its population counterpart because it is so different from what happens in the bulk, where there is a considerable amount of deformation between sample and population eigenvalues. As it turns out, the QuEST function automatically refrains from deforming the arrow eigenvalue, as demonstrated by the following proposition.

Proposition 7.4. *Under Assumptions* 3.1, 3.2.*a–c and* 3.2.*f*, $q_{n,p}^{p}(\boldsymbol{\tau}_n) \sim \tau_{n,p}$.

The similarity between Proposition 7.3 and Proposition 7.4 gives reassurance about the ability of the QuEST function to detect the location of the arrow.

7.4. Shrinking the arrow eigenvalue

Next, we turn to the optimal shrinkage formula. It is not trivial to define what "optimal" means for the arrow because Theorem 3.1 does not take into account finite-rank perturbations. It is necessary to go back to the finite-sample framework of Section 2. In finite samples, the optimal nonlinear shrinkage formula is given by the following lemma.

Lemma 7.1. *Under Assumptions 2.1–2.3, the covariance matrix estimator in the rotation-equivariant class of Definition 2.1 that minimizes Stein's loss in finite samples is*

$$S_n^{\mathrm{FS}} := U_n D_n^{\mathrm{FS}} U_n', \qquad \text{where } D_n^{\mathrm{FS}} := \mathsf{Diag}\left(\frac{1}{u_{n,1}' \Sigma_n^{-1} u_{n,1}}, \dots, \frac{1}{u_{n,p}' \Sigma_n^{-1} u_{n,p}} \right). \tag{7.1}$$

This finite-sample optimal estimator cannot be constructed in practice because it depends on the inverse of the population covariance matrix. But it shows that the optimal nonlinear shrinkage of the sample eigenvalues transforms $\lambda_{n,p}$ into $1/u_{n,p}' \Sigma_n^{-1} u_{n,p}$. The limit of this quantity in an arrow model under large-dimensional asymptotics is given by the following proposition.

Proposition 7.5. *Under Assumptions 3.1, 3.2.a–c, 3.2.f and 3.3,*

$$\frac{1}{u_{n,p}' \Sigma_n^{-1} u_{n,p}} \overset{\text{a.s.}}{\sim} \frac{\tau_{n,p}}{1+c}. \tag{7.2}$$

This is also a surprising result: given that the sample arrow eigenvalue is close to the population arrow eigenvalue, one might have expected that the optimally shrunk arrow eigenvalue would be close to it also. But it is in fact smaller by a factor $1+c$. This poses a stern test for our proposed covariance matrix estimator: will the optimal nonlinear shrinkage formula recognize the need to apply a divisor, and if so will it find the correct arrow shrinkage coefficient of $1+c$? The next proposition answers both questions in the affirmative.

Proposition 7.6. *Under Assumptions 3.1, 3.2.a–c, 3.2.f and 3.3,*

$$\frac{\lambda_{n,p}}{1 - \frac{p}{n} - 2\frac{p}{n}\lambda_{n,p}\,\mathsf{Re}\left[\breve{m}_{n,p}^{\tau_n}(\lambda_{n,p})\right]} \overset{\text{a.s.}}{\sim} \frac{\tau_{n,p}}{1+c}. \tag{7.3}$$

The similarity between Proposition 7.5 and Proposition 7.6 gives reassurance about the ability of the nonlinear shrinkage formula in Corollary 3.1.a to shrink the arrow optimally.

7.5. Wrap-up

What happens at the arrow level has vanishingly small impact on what happens in the bulk because: (i) the gap between the group of bulk eigenvalues and the arrow eigenvalue widens

up to infinity; (ii) the magnitude of the influence between eigenvalues is controlled by the mathematical structure of the Stieltjes transform, making it *inversely proportional* to the distance between them; and (iii) the proportion of eigenvalues in the bulk converges to one.

The bottom line is that the presence of an arrow should not pose any special challenge to our approach for the following reasons:

- spectral decomposition separates the arrow from the bulk due to its signature variability;
- the QuEST function recognizes that sample and population arrow eigenvalues are close;
- our nonlinear shrinkage formula correctly divides the arrow sample eigenvalue by $1 + c$;
- and nonlinear shrinkage of bulk sample eigenvalues remains largely unaffected.

This analysis does not pretend to tie up all the loose ends, but we believe that the accumulated mathematical evidence is sufficient to alleviate potential concerns on this front. To get to the bottom of this matter would require a comprehensive overhaul of the underlying probability theory literature, which obviously lies beyond the scope of the present paper. The theoretical results presented in this section lay the foundation for more in-depth studies of the arrow model, and go a long way towards explaining why our nonlinear shrinkage estimator performs well numerically in the arrow model simulated in Section 8.

8. 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 [42] 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 (2.2). The article by Lin and Perlman [30] 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. Nonetheless, we do include the sample covariance matrix and the linear shrinkage estimator of Ledoit and Wolf [26]; we also include the minimax estimator of Dey and Srinivasan [12], Theorem 3.1.

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

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

Thus $\mathrm{PRIAL}(S_n^{\mathrm{ST}}, S_n^{\mathrm{ST}}) = 0\%$ and $\mathrm{PRIAL}(S_n^{\mathrm{ST}}, \Sigma_n) = 100\%$ by construction. The quantity that we report is $\mathrm{PRIAL}(S_n^{\mathrm{ST}}, \widehat{\Sigma}_n)$, where the empirical risks of S_n^{ST} and $\widehat{\Sigma}_n$ are computed as averages across 1000 Monte Carlo simulations.

Unless stated otherwise, the ith population eigenvalue is equal to $\tau_{n,i} := H^{-1}((i - 0.5)/p)$ $(i = 1, \ldots, p)$, 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 $W \sim \mathrm{Beta}(2, 5)$. This distribution is right-skewed, meaning that there are a lot of small eigenvalues and a few large ones, which is representative of many practically relevant situations; see the solid line in Figure 4 below. In this scenario, the PRIAL of our new nonlinear shrinkage estimator relative to Stein's is 43%.

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. The minimax estimator and the sample covariance matrix fail to beat Stein's estimator. Both linear and nonlinear shrinkage improve over Stein; the improvement is strong across the board, and stronger in small-to-medium dimensions.

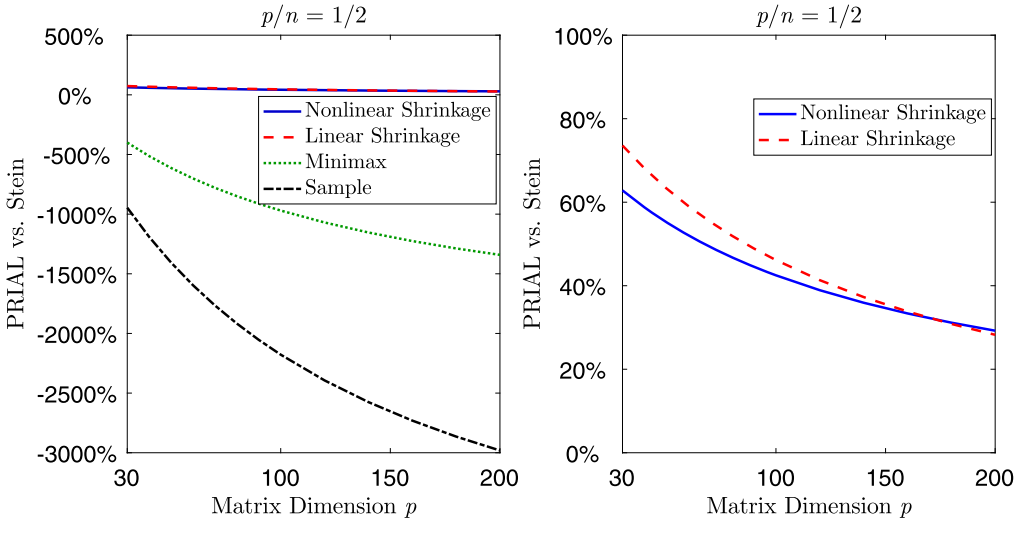

Figure 1. Evolution of the PRIAL of various estimators relative to Stein's estimator as matrix dimension and sample size go to infinity together. The left panel shows all the results, whereas the right panel zooms in on positive improvements for better visual clarity.

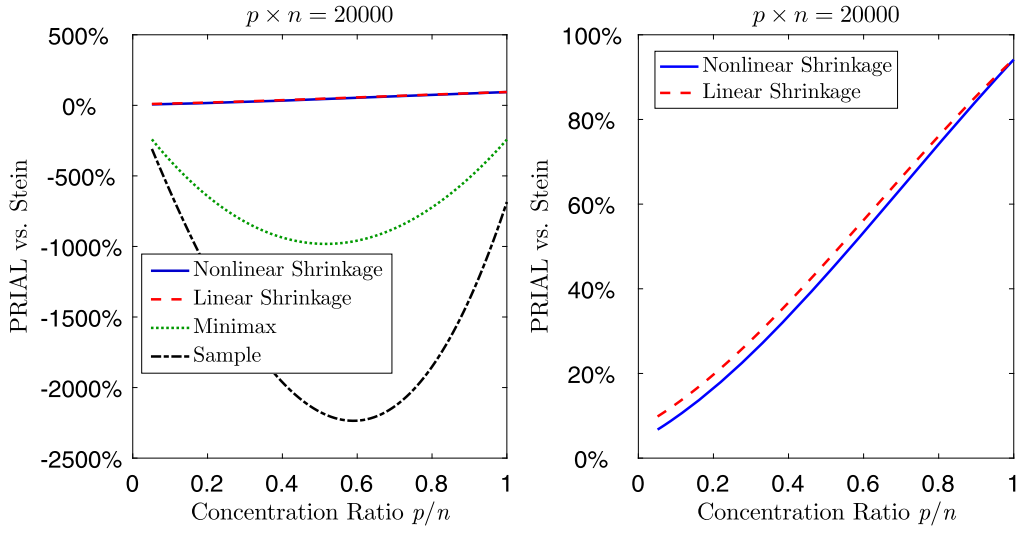

Figure 2. PRIAL of four different estimators relative to Stein's estimator as a function of the concentration ratio p/n. The left panel shows all the results, whereas the right panel zooms in on positive improvements for better visual clarity.

Concentration

Second, we vary the concentration (ratio) from $p/n = 0.05$ to $p/n = 1.0$ while keeping the product $p \times n$ constant at the value 20 000. The results are displayed in Figure 2. Once again, minimax performs better than the sample covariance matrix, yet worse than Stein's estimator. Both linear and nonlinear shrinkage improve over Stein; the improvement is good across the board and stronger when the matrix dimension is close to the sample size. In particular, nonlinear shrinkage can handle the case $p/n = 1$ even though it is not covered by the mathematical treatment; see Remark 6.1.

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. The minimax estimator and the sample covariance matrix again fail to beat Stein's estimator. The improvement delivered by nonlinear shrinkage is always strictly positive and increases as the population covariance matrix becomes better conditioned. Linear shrinkage always beats the sample covariance matrix but has otherwise mixed results, possibly due to the fact that it is optimized with respect to the Frobenius loss instead of Stein's loss.

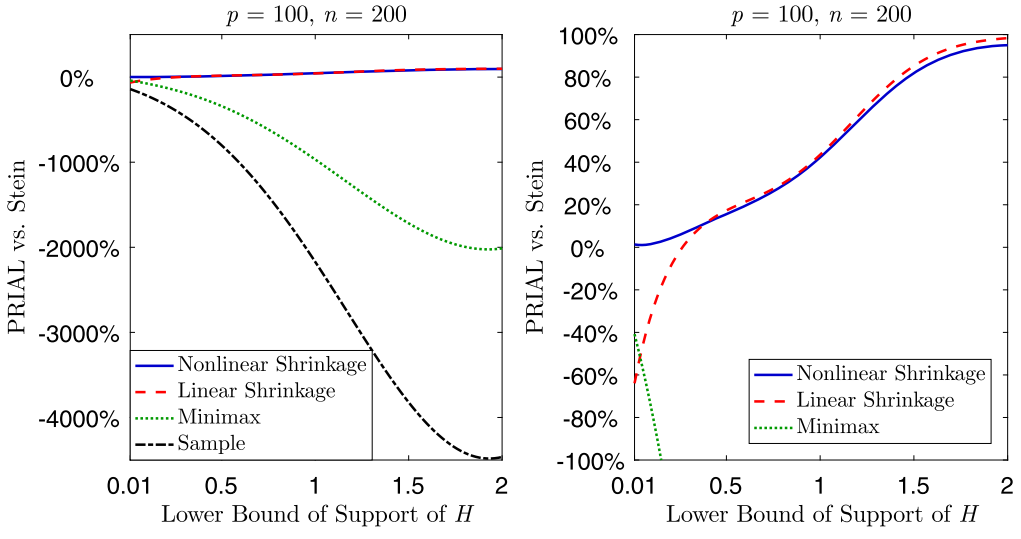

Figure 3. PRIAL of four estimators relative to Stein's estimator across various condition numbers. The left panel shows all the results, whereas the right panel zooms in on the range of PRIALs between ±100% for better visual clarity.

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. The results are presented in Figure 5. To preserve the clarity of the picture, we only report the PRIAL of the nonlinear shrinkage estimator; but the other results are in line with Figure 1. 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.

Clustered eigenvalues

Fifth, we consider a different type of distribution for the population eigenvalues: a *discrete* distribution. More specifically, we assume that the population covariance matrix has 20% of its eigenvalues equal to 1, 40% equal to 3 and 40% equal to 10. This is a particularly interesting and difficult example introduced and analyzed in detail by Bai and Silverstein [2]; in particular, it produces highly nonlinear patterns. As in Figures 1 and 4, 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 6. Nonlinear shrinkage improves over Stein for all dimensions, though not by much. The other estimators are worse than Stein for all dimensions. Linear shrinkage is at a disadvantage in this setup due to the highly nonlinear nature of the optimal shrinkage transformation.

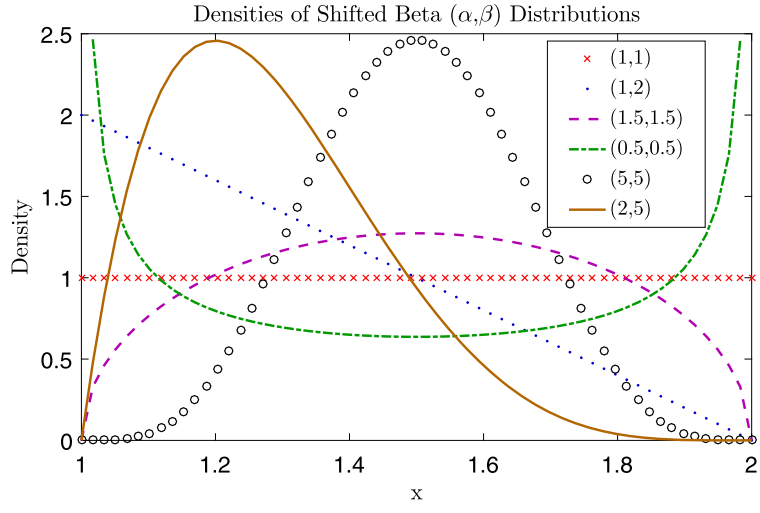

Figure 4. Densities of various shifted Beta distributions. Note that the density of Beta(β, α) is just the mirror image (around the mid point of the support) of the density of Beta(α, β).

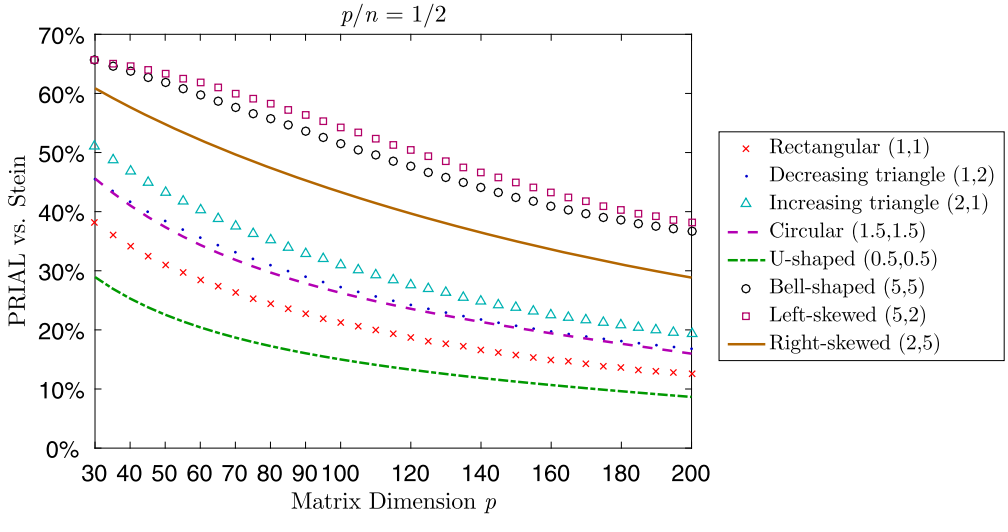

Figure 5. PRIAL of the nonlinear shrinkage estimator relative to Stein's estimator for various shapes of the population spectral distribution.

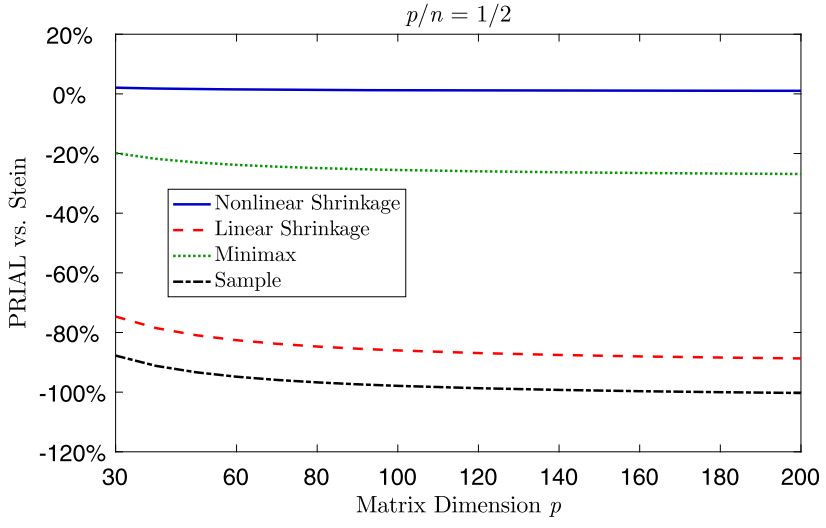

Figure 6. Evolution of the PRIAL of various estimators relative to Stein's estimator as matrix dimension and sample size go to infinity together. The limiting population spectral distribution is discrete.

Non-normality

Sixth, 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 1. One can see that the results in the normal case carry over qualitatively to the non-normal cases.

Singular case with fixed concentration ratio

Seventh, 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

Table 1. PRIAL for different distributions of the variates

Distribution	Nonlinear	Linear	Minimax	Sample
Normal	43%	46%	−983%	−2210%
Bernoulli	42%	43%	−1020%	−2307%
Laplace	44%	52%	−889%	−1980%

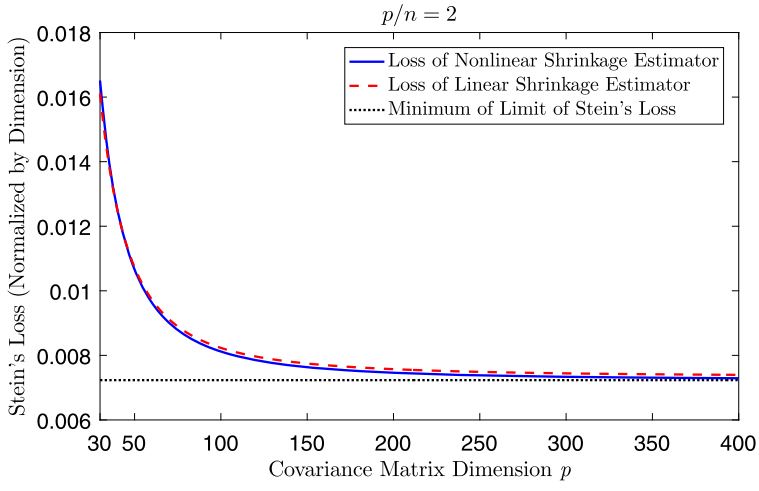

Figure 7. Stein's loss for the linear and nonlinear shrinkage estimators when dimension exceeds sample size. The benchmark is the minimum of the limit of Stein's loss among rotation-equivariant estimators.

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 1000 Monte Carlo simulations for our nonlinear shrinkage estimator and for linear shrinkage is displayed in Figure 7. These results confirm that our nonlinear shrinkage estimator minimizes Stein's loss asymptotically even in the difficult case where variables outnumber observations. The loss of the linear shrinkage estimator is slightly higher. Due to the fact that $p > n$, Stein's and the minimax estimator are not defined, and the loss of the sample covariance matrix is not defined either.

Singular case with fixed matrix dimension

In order to further study the singular case, we fix the matrix dimension p at a high number, in this case $p = 1000$, and let the sample size n vary from $n = 100$ to $n = 1000$.[2] We take the same distribution for H as in the baseline scenario and simulate Gaussian variates. The concentration ratio varies from $c = 10$ to $c = 1$. The average loss across 1000 Monte Carlo simulations for our nonlinear shrinkage estimator and for linear shrinkage is displayed in Figure 8. The loss of the linear shrinkage estimator is higher than the loss of our nonlinear shrinkage estimator, especially for large concentrations. Since $p \geq n$, Stein's estimator and the minimax estimator are not defined, and the loss of the sample covariance matrix is not defined either.

[2]We thank an anonymous referee for suggesting this numerical experiment.

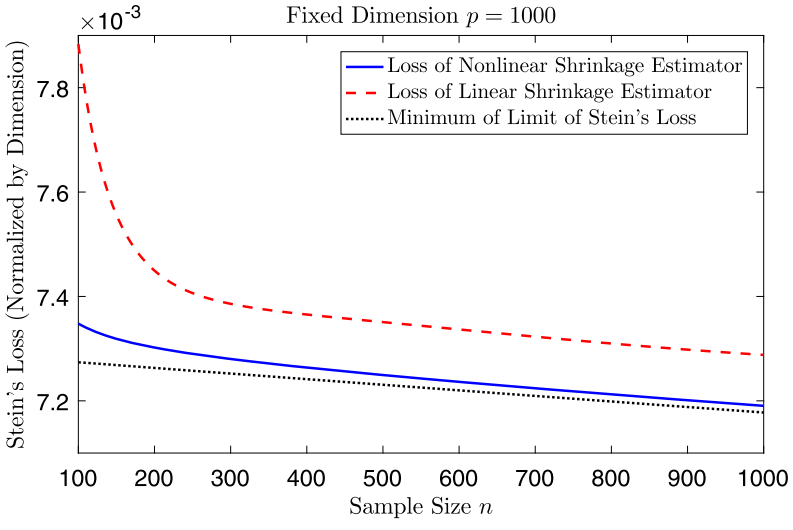

Figure 8. Stein's loss for the linear and nonlinear shrinkage estimators in the singular case with fixed dimension. The benchmark is the minimum of the limit of Stein's loss among rotation-equivariant estimators.

Arrow model

Finally, we study the performance of our nonlinear shrinkage estimator in the case where the largest population eigenvalue is of order n, in violation of Assumption 3.2.d. Inspired by a factor model where all pairs of variables have 50% correlation and all variables have unit standard deviation, and by the arrow model defined by Assumption 3.2.f, we set $\tau_{n,p}$ equal to $1 + 0.5(p - 1)$. The other eigenvalues are set as per the baseline scenario. Thus, $\boldsymbol{\tau}_n := (H^{-1}(0.5/(p-1)), \ldots, H^{-1}((p-1.5)/(p-1)), 1 + 0.5(p-1))'$, 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 9. Our nonlinear shrinkage estimator still convincingly dominates Stein's estimator, although Assumption 3.2.d is violated; none of the other estimators can beat Stein.

Overall assessment

We have conducted an extensive set of Monte Carlo simulatons. To start with the obvious, both the sample covariance matrix and the minimax estimator perform universally worse than the reference estimator of Stein [41,42], often with PRIALs in the -1000% zone.

More intriguing is the competition between Stein's estimator and linear shrinkage. In theory, this setup should favor Stein because linear shrinkage is so much simpler (by glancing over nonlinearities) and minimizes the Frobenius loss instead of Stein's loss. In practice, linear shrinkage

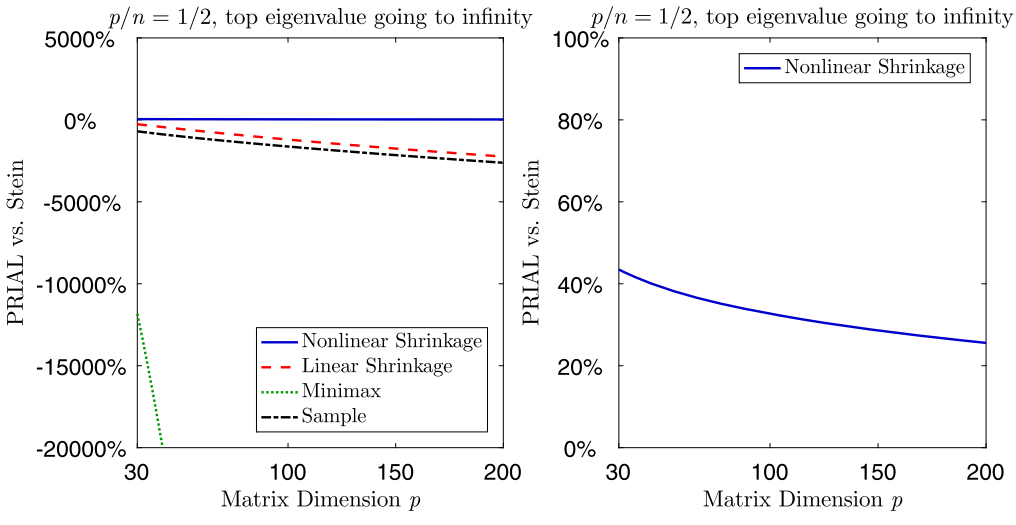

Figure 9. PRIAL of the new nonlinear shrinkage estimator relative to Stein's estimator when the top eigenvalue goes to infinity. The left panel shows all the results above $-20\,000\%$, whereas the right panel zooms in on the results above 0% for better visual clarity.

still beats Stein across a wide variety of situations, often by a substantial margin; for example, see Figures 1 and 2. Conversely, Stein's estimator improves over linear shrinkage in other situations where nonlinear effects are more prominent; for example, see Figures 6 and 9. Thus, we witness the emergence of two different regimes.

How does nonlinear shrinkage perform? In the regime where linear shrinkage soundly beats Stein, nonlinear shrinkage also improves over Stein by a similarly substantial margin. In the other regime where Stein beats linear shrinkage, nonlinear shrinkage always dominates Stein. Thus, nonlinear shrinkage can be said to combine the best of Stein and linear shrinkage.

9. Empirical application

The goal of this section is to examine the out-of-sample properties of Markowitz portfolios based on various covariance matrix estimators.

9.1. Data and general portfolio-formation rules

We download daily data from the Center for Research in Security Prices (CRSP) starting in 01/01/1980 and ending in 12/31/2015. For simplicity, we adopt the common convention that 21 consecutive trading days constitute one 'month'. The out-of-sample period ranges from

01/08/1986 through 12/31/2015, resulting in a total of 360 months (or 7560 days). All portfolios are updated monthly.[3] We denote the investment dates by $h = 1, \ldots, 360$. At any investment date h, a covariance matrix is estimated using the most recent $n = 252$ daily returns, which roughly corresponds to using one year of past data.

We consider the following portfolio sizes: $p \in \{50, 100, 150, 200, 250\}$. For a given combination (h, p), the investment universe is obtained as follows. We find the set of stocks that have a complete return history over the most recent $n = 252$ days as well as a complete return 'future' over the next 21 days.[4] We then look for possible pairs of highly correlated stocks, that is, pairs of stocks that returns with a sample correlation exceeding 0.95 over the past 252 days. With such pairs, if they should exist, we remove the stock with the lower volume of the two on investment date h.[5] Of the remaining set of stocks, we then pick the largest p stocks (as measured by their market capitalization on investment date h) as our investment universe. In this way, the investment universe changes slowly from one investment date to the next.

9.2. Global minimum variance portfolio

We consider the problem of estimating the global minimum variance (GMV) portfolio, in the absence of short-sales constraints. The problem is formulated as

$$\min_{w} w'\Sigma w \tag{9.1}$$

$$\text{subject to} \quad w'\mathbb{1} = 1, \tag{9.2}$$

where $\mathbb{1}$ denotes a vector of ones of dimension $p \times 1$. It has the analytical solution

$$w = \frac{\Sigma^{-1}\mathbb{1}}{\mathbb{1}'\Sigma^{-1}\mathbb{1}}. \tag{9.3}$$

The natural strategy in practice is to replace the unknown Σ by an estimator $\widehat{\Sigma}$ in formula (9.3), yielding a feasible portfolio

$$\widehat{w} := \frac{\widehat{\Sigma}^{-1}\mathbb{1}}{\mathbb{1}'\widehat{\Sigma}^{-1}\mathbb{1}}. \tag{9.4}$$

Estimating the GMV portfolio is a clean problem in terms of evaluating the quality of a covariance matrix estimator, since it abstracts from having to estimate the vector of expected returns at the same time. In addition, researchers have established that estimated GMV portfolios have

[3]Monthly updating is common practice to avoid an unreasonable amount of turnover and thus transaction costs. During a month, from one day to the next, we hold number of shares fixed rather than portfolio weights; in this way, there are no transactions at all during a month.

[4]The latter, forward-looking restriction is not a feasible one in real life but is commonly applied in the related finance literature on the out-of-sample evaluation of portfolios.

[5]The reason is that we do not want to include highly similar stocks.

desirable out-of-sample properties not only in terms of risk but also in terms of reward-to-risk (that is, in terms of the information ratio); for example, see Haugen and Baker [17], Jagannathan and Ma [19], and Nielsen and Aylursubramanian [35]. As a result, such portfolios have become an addition to the large array of products sold by the mutual-fund industry. The following six portfolios are included in the study.

$1/N$**:** the equal-weighted portfolio. This portfolio is a standard benchmark and has been promoted by DeMiguel, Garlappi and Uppal [11], among others. This portfolio can actually seen as a special case of portfolio (9.4), where the 'estimator' $\widehat{\Sigma}$ is simply the identity matrix.
Sample: the portfolio (9.4), where the estimator $\widehat{\Sigma}$ is the sample covariance matrix.
Stein: the portfolio (9.4), where the estimator $\widehat{\Sigma}$ is Stein's estimator.
Minimax: the portfolio (9.4), where the estimator $\widehat{\Sigma}$ is from Dey and Srinivasan [12].
Lin: the portfolio (9.4), where the estimator $\widehat{\Sigma}$ is the estimator of Ledoit and Wolf [26].
NonLin: the portfolio (9.4), where the estimator $\widehat{\Sigma}$ is the estimator of Theorem 5.2.

We report the following three out-of-sample performance measures for each scenario.

AV: We compute the average of the 7560 out-of-sample log returns and then multiply by 252 in order to annualize.
SD: We compute the standard deviation of the 7560 out-of-sample log returns and then multiply by $\sqrt{252}$ in order to annualize.
IR: We compute the (annualized) information ratio as the ratio AV/SD.[6]

Our stance is that in the context of the GMV portfolio, the most important performance measure is the out-of-sample standard deviation, SD. The true (but unfeasible) GMV portfolio is given by (9.3). It is designed to minimize the variance (and thus the standard deviation) rather than to maximize the expected return or the information ratio. Therefore, any portfolio that implements the GMV portfolio should be primarily evaluated by how successfully it achieves this goal. A high out-of-sample average return, AV, and a high out-of-sample information ratio, IR, are naturally also desirable, but should be considered of secondary importance from the point of view of evaluating the quality of a covariance matrix estimator.

The results are presented in Table 2 and can be summarized as follows; unless stated otherwise, the findings are with respect to the standard deviation as performance measure.

- All 'sophisticated' portfolios outperform the 'naïve' $1/N$ portfolio for $p \leq 200$. But for $p = 250$, Sample, Stein, and Minimax break down and underperform $1/N$; on the other hand, Lin and NonLin continue to outperform $1/N$.
- NonLin is uniformly best. For $p \leq 200$, Stein is second-best and Lin is third-best; on the other hand, for $p = 250$, Lin is second-best.
- In terms of the information ratio, NonLin is best followed by Lin and Stein.

[6]This version of the information ratio, which simply uses zero as the benchmark, is widely used in the mutual fund industry.

Table 2. Annualized performance measures (in percent) for various estimators of the GMV portfolio. AV stands for average; SD stands for standard deviation; and IR stands for information ratio. All measures are based on 7560 daily out-of-sample returns from 01/08/1986 through 12/31/2015. In the rows labeled SD, the lowest number appears in **bold face**

	Period: 01/08/1986–12/31/2015					
	1/N	Sample	Stein	Minimax	Lin	NonLin
$p = 50$						
AV	11.87	9.16	9.32	9.32	9.34	9.28
SD	22.78	15.06	14.61	14.71	14.63	**14.58**
IR	0.52	0.61	0.64	0.64	0.64	0.64
$p = 100$						
AV	12.10	8.52	9.38	9.08	9.20	9.39
SD	21.56	14.69	13.06	13.46	13.33	**13.01**
IR	0.56	0.58	0.72	0.67	0.69	0.72
$p = 150$						
AV	12.57	9.84	9.29	9.49	9.41	9.36
SD	21.00	15.64	12.27	12.98	12.67	**12.16**
IR	0.60	0.63	0.76	0.73	0.74	0.77
$p = 200$						
AV	12.67	9.71	9.56	9.91	10.13	9.70
SD	20.57	19.56	11.80	13.27	12.12	**11.49**
IR	0.61	0.49	0.81	0.75	0.84	0.84
$p = 250$						
AV	13.15	43.24	25.52	20.03	10.63	9.57
SD	20.24	245.50	82.91	52.49	11.72	**11.00**
IR	0.65	0.18	0.31	0.38	0.91	0.88

10. 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 [41,42] 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 covariance matrix estimator that is not positive-semidefinite). 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 a 'limiting shrinkage function' – provide a more formal justification to estimators based on the Frobenius loss proposed by Ledoit and Wolf [27,28].

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 [27,28]. 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 Monte Carlo simulations.

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

An in-depth exploration of what we call the "arrow model" – where the largest population eigenvalue goes to infinity at the same rate as the matrix dimension – and of its implications for covariance matrix estimation are also left as a fruitful avenue for future research.

Acknowledgements

We thank an associate editor and two anonymous referees for helpful comments that have greatly enhanced the exposition of the paper. We also thank seminar participants at the Department of Statistics at Stanford University, and especially Bala Rajaratnam, for feedback on an earlier version of this paper. All remaining errors are ours.

Supplementary Material

Supplement: Proofs of mathematical results (DOI: 10.3150/17-BEJ979SUPP; .pdf). This supplement collects the proofs of all mathematical results.

References

[1] Bai, Z. and Silverstein, J.W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed. *Springer Series in Statistics*. New York: Springer. MR2567175

[2] Bai, Z.D. and Silverstein, J.W. (1998). No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *Ann. Probab.* **26** 316–345. MR1617051

[3] Bai, Z.D. and Silverstein, J.W. (1999). Exact separation of eigenvalues of large-dimensional sample covariance matrices. *Ann. Probab.* **27** 1536–1555. MR1733159

[4] Bai, Z.D. and Silverstein, J.W. (2004). CLT for linear spectral statistics of large-dimensional sample covariance matrices. *Ann. Probab.* **32** 553–605. MR2040792

[5] Bai, Z.D., Silverstein, J.W. and Yin, Y.Q. (1988). A note on the largest eigenvalue of a large-dimensional sample covariance matrix. *J. Multivariate Anal.* **26** 166–168. MR0963829

[6] Baik, J., Ben Arous, G. and Péché, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.* **33** 1643–1697. MR2165575

[7] Baik, J. and Silverstein, J.W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivariate Anal.* **97** 1382–1408. MR2279680

[8] Bickel, P.J. and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

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

[10] Daniels, M.J. and Kass, R.E. (2001). Shrinkage estimators for covariance matrices. *Biometrics* **57** 1173–1184. MR1950425

[11] DeMiguel, V., Garlappi, L. and Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the $1/N$ portfolio strategy? *Rev. Financ. Stud.* **22** 1915–1953.

[12] Dey, D.K. and Srinivasan, C. (1985). Estimation of a covariance matrix under Stein's loss. *Ann. Statist.* **13** 1581–1591. MR0811511

[13] Donoho, D.L., Gavish, M. and Johnstone, I.M. (2014). Optimal shrinkage of eigenvalues in the spiked covariance model. arXiv:1311.0851v2.

[14] El Karoui, N. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Ann. Statist.* **36** 2757–2790. MR2485012

[15] Gill, P.E., Murray, W. and Saunders, M.A. (2002). SNOPT: An SQP algorithm for large-scale constrained optimization. *SIAM J. Optim.* **12** 979–1006. MR1922505

[16] Haff, L.R. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix. *Ann. Statist.* **8** 586–597. MR0568722

[17] Haugen, R.A. and Baker, N.L. (1991). The efficient market inefficiency of capitalization-weighted stock portfolios. *J. Portf. Manag.* **17** 35–40.

[18] Henrici, P. (1988). *Applied and Computational Complex Analysis* **1**. New York: Wiley.

[19] Jagannathan, R. and Ma, T. (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *J. Finance* **54** 1651–1684.

[20] James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I* 361–379. Berkeley, CA: Univ. California Press. MR0133191

[21] Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **186** 453–461. MR0017504

[22] Johnstone, I.M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961

[23] Khare, K., Oh, S.-Y. and Rajaratnam, B. (2015). A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 803–825. MR3382598

[24] Kullback, S. and Leibler, R.A. (1951). On information and sufficiency. *Ann. Math. Stat.* **22** 79–86. MR0039968
[25] Ledoit, O. and Péché, S. (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probab. Theory Related Fields* **151** 233–264. MR2834718
[26] Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.* **88** 365–411. MR2026339
[27] Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Ann. Statist.* **40** 1024–1060. MR2985942
[28] Ledoit, O. and Wolf, M. (2015). Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions. *J. Multivariate Anal.* **139** 360–384.
[29] Ledoit, O. and Wolf, M. (2017). Supplement to "Optimal estimation of a large-dimensional covariance matrix under Stein's loss". DOI:10.3150/17-BEJ979SUPP.
[30] Lin, S.P. and Perlman, M.D. (1985). A Monte Carlo comparison of four estimators of a covariance matrix. In *Multivariate Analysis VI* (*Pittsburgh*, *Pa.*, 1983) 411–429. Amsterdam: North-Holland. MR0822310
[31] Marčenko, V.A. and Pastur, L.A. (1967). Distribution of eigenvalues for some sets of random matrices. *Sb. Math.* **1** 457–483.
[32] Mestre, X. (2008). Improved estimation of eigenvalues and eigenvectors of covariance matrices using their sample estimates. *IEEE Trans. Inform. Theory* **54** 5113–5129. MR2589886
[33] Moakher, M. and Batchelor, P.G. (2006). Symmetric positive-definite matrices: From geometry to applications and visualization. In *Visualization and Processing of Tensor Fields*. *Math. Vis.* 285–298, 452. Berlin: Springer. MR2210524
[34] Muirhead, R.J. (1982). *Aspects of Multivariate Statistical Theory*. *Wiley Series in Probability and Mathematical Statistics*. New York: John Wiley & Sons, Inc. MR0652932
[35] Nielsen, F. and Aylursubramanian, R. (2008). Far From the Madding Crowd – Volatility Efficient Indices. Research Insights, MSCI Barra.
[36] Rajaratnam, B. and Vincenzi, D. (2016). A theoretical study of Stein's covariance estimator. *Biometrika* **103** 653–666. MR3551790
[37] Silverstein, J.W. (1995). Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. *J. Multivariate Anal.* **55** 331–339. MR1370408
[38] Silverstein, J.W. and Bai, Z.D. (1995). On the empirical distribution of eigenvalues of a class of large-dimensional random matrices. *J. Multivariate Anal.* **54** 175–192. MR1345534
[39] Silverstein, J.W. and Choi, S.-I. (1995). Analysis of the limiting spectral distribution of large-dimensional random matrices. *J. Multivariate Anal.* **54** 295–309. MR1345541
[40] Stein, C. (1956). Some problems in multivariate analysis, Part I. Technical Report No. 6, Department of Statistics, Stanford Univ.
[41] Stein, C. (1975). Estimation of a covariance matrix. Rietz lecture, 39th Annual Meeting IMS. Atlanta, Georgia.
[42] Stein, C. (1986). Lectures on the theory of estimation of many parameters. *J. Math. Sci.* **34** 1373–1403.
[43] Tsukuma, H. (2005). Estimating the inverse matrix of scale parameters in an elliptically contoured distribution. *J. Japan Statist. Soc.* **35** 21–39. MR2183498
[44] Won, J.-H., Lim, J., Kim, S.-J. and Rajaratnam, B. (2013). Condition-number-regularized covariance estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 427–450. MR3065474
[45] Yin, Y.Q., Bai, Z.D. and Krishnaiah, P.R. (1988). On the limit of the largest eigenvalue of the large-dimensional sample covariance matrix. *Probab. Theory Related Fields* **78** 509–521. MR0950344

Received April 2016 and revised July 2017