Year: 2022

Quadratic shrinkage for large covariance matrices

Ledoit, Olivier ; Wolf, Michael

Abstract: This paper constructs a new estimator for large covariance matrices by drawing a bridge between the classic (Stein (1975)) estimator in finite samples and recent progress under large-dimensional asymptotics. The estimator keeps the eigenvectors of the sample covariance matrix and applies shrinkage to the inverse sample eigenvalues. The corresponding formula is quadratic: it has two shrinkage targets weighted by quadratic functions of the concentration (that is, matrix dimension divided by sample size). The first target dominates mid-level concentrations and the second one higher levels. This extra degree of freedom enables us to outperform linear shrinkage when the optimal shrinkage is not linear, which is the general case. Both of our targets are based on what we term the "Stein shrinker", a local attraction operator that pulls sample covariance matrix eigenvalues towards their nearest neighbors, but whose force diminishes with distance (like gravitation). We prove that no cubic or higher-order nonlinearities beat quadratic with respect to Frobenius loss under large-dimensional asymptotics. Non-normality and the case where the matrix dimension exceeds the sample size are accommodated. Monte Carlo simulations confirm state-of-the-art performance in terms of accuracy, speed, and scalability.

DOI: https://doi.org/10.3150/20-bej1315

Originally published at:
Ledoit, Olivier; Wolf, Michael (2022). Quadratic shrinkage for large covariance matrices. Bernoulli, 28(3):1519-1547.
DOI: https://doi.org/10.3150/20-bej1315

Bernoulli **28**(3), 2022, 1519–1547
https://doi.org/10.3150/20-BEJ1315

Quadratic shrinkage for large covariance matrices

OLIVIER LEDOIT[1,2,a,c] and MICHAEL WOLF[1,b]

[1]*Department of Economics, University of Zurich, 8032 Zurich, Switzerland.* [a]*olivier.ledoit@econ.uzh.ch,* [b]*michael.wolf@econ.uzh.ch*
[2]*AlphaCrest Capital Management, New York, NY 10036, USA.* [c]*olivier.ledoit@alphacrestcapital.com*

This paper constructs a new estimator for large covariance matrices by drawing a bridge between the classic (Stein (1975)) estimator in finite samples and recent progress under large-dimensional asymptotics. The estimator keeps the eigenvectors of the sample covariance matrix and applies shrinkage to the *inverse* sample eigenvalues. The corresponding formula is *quadratic*: it has two shrinkage targets weighted by quadratic functions of the concentration (that is, matrix dimension divided by sample size). The first target dominates mid-level concentrations and the second one higher levels. This extra degree of freedom enables us to outperform linear shrinkage when the optimal shrinkage is not linear, which is the general case. Both of our targets are based on what we term the "Stein shrinker", a local attraction operator that pulls sample covariance matrix eigenvalues towards their nearest neighbors, but whose force diminishes with distance (like gravitation). We prove that no cubic or higher-order nonlinearities beat quadratic with respect to Frobenius loss under large-dimensional asymptotics. Non-normality and the case where the matrix dimension exceeds the sample size are accommodated. Monte Carlo simulations confirm state-of-the-art performance in terms of accuracy, speed, and scalability.

Keywords: Inverse shrinkage; kernel estimation; large-dimensional asymptotics; signal amplitude; Stein shrinkage

1. Introduction

The covariance matrix is, arguably, the second most important object in all of statistics. It has long been known — by theoreticians and practitioners alike — that the sample covariance matrix suffers from the curse of dimensionality. This curse is most obvious when the matrix dimension exceeds the sample size — in which case the sample covariance matrix is singular — but is pervasive unless the matrix dimension is *negligible* with respect to the sample size.

Efforts to robustify covariance matrix estimation against large dimensions can be broadly divided into two generations. First, the 20th century, characterized by (1.a) finite-sample mathematics, (1.b) the normality assumption, and (1.c) matrix dimensions below the sample size. Second, the 21st century, characterized by (2.a) large-dimensional asymptotics, (2.b) relaxation of the normality assumption, and (2.c) matrix dimensions below or above the sample size. At the risk of over-simplification, the second generation can be summarized as giving first-generation (finite-sample) ideas a major upgrade thanks to the powerful mathematics of large-dimensional asymptotics. For example, Ledoit and Wolf (2004) apply linear shrinkage to the sample covariance matrix in the spirit of James and Stein (1961), whereas Bodnar, Gupta and Parolya (2016) apply linear shrinkage to the sample precision matrix (inverse of the sample covariance matrix) as Haff (1979) did. The tests for identity and sphericity of the covariance matrix proposed by Ledoit and Wolf (2002) and Chen, Zhang and Zhong (2010) are direct heirs of the ones of John (1971) and Nagao (1973).

The crowning achievement of the first generation of research on robustifying covariance matrix estimation against the curse of dimensionality is without doubt the nonlinear shrinkage estimator of Stein (1977, 1975, 1986). A string of Monte Carlo simulations starting with Lin and Perlman (1985)

1350-7265 © 2022 ISI/BS

have found it remarkably accurate, especially when the cross-sectional distribution of covariance matrix eigenvalues (a.k.a. principal components) is not smooth but clustered, a difficult case to handle. According to Rajaratnam and Vincenzi (2016a), “Stein’s covariance estimator is considered a gold standard in the literature”. Nevertheless, the Stein shrinkage estimator has not been given its rightful ‘large-dimensional-asymptotic upgrade’ yet. Doing so is the objective of the present paper.

We start from a solid foundation in finite samples by reinterpreting Stein’s highly nonlinear (and not immediately intuitive) formula as just linear shrinkage in inverse-eigenvalues space. This gives much greater clarity and insight into what is really happening under the hood. The linear shrinkage intensity could not be simpler: it is the “concentration (ratio)”, defined as the ratio of matrix dimension to sample size, a standard measure of the severity of the curse of dimensionality. The higher the concentration, the more the eigenvalues need to be shrunk away from the observed ones. This is called “shrinkage” because the cross-sectional dispersion of the eigenvalues goes down, as they are attracted to one another. What is more interesting is that the shrinkage target is not always the same: it varies depending on relative position with respect to surrounding sample eigenvalues. An eigenvalue that lies slightly above (below) a concentrated cluster of the other eigenvalues is attracted downwards (upwards), and the intensity of this attraction vanishes as the distance increases. Stein’s results, reinterpreted in this light, provide a well-defined targeting function that captures this important phenomenon, and we call this function the “Stein shrinker”. It will be the central object throughout the paper.

The only problem with the naïve Stein shrinker is that it explodes when two sample eigenvalues get too close to each other. This is where large-dimensional asymptotics comes into play. We show that smoothing the Stein shrinker provides a covariance matrix estimator that is optimal with respect to Stein’s loss under large-dimensional asymptotics. The smoothing parameter must vanish asymptotically as the matrix dimension and the sample size go to infinity together, but not too fast. Note that Stein himself, even though he had essentially the same formula (up to smoothing), could not formally prove optimality in finite samples. Note also that Stein (1986) explicitly acknowlegded both the need for injecting some type of smoothing (*ex-post* through his so-called isotonization algorithm), and also the relevance of large-dimensional asymptotics (cf. his Theorem 1). Therefore, the “smoothed Stein shrinker”, as we call it, is just a reorganization of the fundamental ingredients that were already embedded in Stein’s original work, but could not be brought to full fruition at that time.

These developments naturally open the door to other loss functions. The first obvious candidate is simply the Inverse Stein’s loss (Tsukuma, 2005): Stein’s loss applied to the precision matrix, instead of the covariance matrix itself. This loss function belongs to a broader class that contains two more loss functions: the Frobenius loss, and the Minimum Variance loss of Engle, Ledoit and Wolf (2019). The former has proven quite popular in a variety of applied fields ranging from macroeconomics (Korniotis, 2008) to brain-computer interface (Vidaurre *et al.*, 2009) to analytical chemistry (Guo *et al.*, 2012), among many others. The latter is ideal whenever the objective is to optimize the reward-to-risk ratio (finance) or signal-to-noise ratio (electrical engineering). The reason why these three loss functions are grouped together is that they lead to the same optimal nonlinear shrinkage formula, both in finite samples and under large-dimensional asymptotics.

To handle these loss functions, the asymptotically optimal solution is to move from *linear* shrinkage in inverse-eigenvalues space to *quadratic* shrinkage. There are now two shrinkage targets: one driven by the smoothed Stein shrinker as before, and the other by its “squared amplitude”. For readers not versed in signal processing, the squared amplitude is basically the square of the Stein shrinker, but with something more added. The extra part is the square of the ‘hidden’ or imaginary component that is the conjugate of the Stein shrinker. This is a basic concept in signal processing that goes back to Gabor (1946), but we review all the necessary details in the main body of the paper. As for the shrinkage intensities themselves, they are split three-ways between the original inverse eigenvalues, the smoothed Stein shrinker, and its squared amplitude. The first dominates small concentrations, the second dominates for mid-level concentrations around the 0.5 mark, and the third dominates for high concentrations tending to 1. All three shrinkage intensities are quadratic functions of the concentration ratio.

To cap it all off, we venture into the realm where the sample covariance matrix is singular. In this case the quadratic weighting scheme delineated above degenerates into putting 100% weight on the squared amplitude of the smoothed Stein shrinker. The null sample eigenvalues need to be kicked out of the computation of the shrinker, so instead they get their own shrinkage formula: a simple function of the concentration ratio and the harmonic mean of the non-null eigenvalues.

Rolling out the 'large-dimensional asymptotic upgrade' generates many obvious contributions with respect to Stein (1975, 1986), listed above. With respect to linear shrinkage, the contribution is also clear because going from linear to quadratic is an easily manageable enhancement that guarantees maximum accuracy even when eigenvalues can be dispersed, clustered, or otherwise unruly. With respect to the other nonlinear shrinkage formulas from Ledoit and Wolf (2015) onwards, there are two key advantages. The first is that the formula obtained here comes from classical statistics instead of from random matrix theory (RMT). All existing ones have been based on a fundamental equation from RMT originally due to Marčenko and Pastur (1967), reformulated by Silverstein (1995), and generalized by Ledoit and Péché (2011). By starting from Stein's (1975) first-generation classic instead, we not only reconnect with a rich body of literature in multivariate statistics, but also inject much-needed understandability. Although many results from RMT have been used for quite a while now by some statisticians, the field is, arguably, still not overly familiar to others. By contrast, it is plain to see, just by visually inspecting the Stein shrinker itself, that eigenvalues are attracted to close-by clusters of other eigenvalues, whereas distant clusters have diminishing influence. We hope that this feature makes the resulting estimator more transparent and user-friendly because opacity usually slows down adoption. The second key advantage is that we manage to reduce mathematical complexity from an infinite degree of nonlinearity to just two degrees (quadratic shrinkage). All this is accomplished without sacrificing accuracy, computational speed, or scalability. In particular, concerning accuracy, we attain the same performance in the large-dimensional limit as the best nonlinear shrinkage formulas based on the fundamental equation of RMT, and (almost) the same performance in finite samples as well.

The remainder of the paper is organized as follows. Section 2 reinterprets the classic first-generation Stein (1986) method as linear shrinkage in inverse-eigenvalues space and introduces what we term the "Stein shrinker". Section 3 shows how to smooth out the explosive discontinuity inside the Stein shrinker and states conditions on the smoothing parameter that guarantee optimality with respect to Stein's loss under large-dimensional asymptotics. Section 4 adapts the formula to the Inverse Stein, Frobenius, and Minimum Variance loss functions by introducing a second shrinkage target based on the squared amplitude of the smoothed Stein shrinker, and by making shrinkage intensities quadratic functions of the concentration ratio. Section 5 shows how to handle the case when the dimension exceeds the sample size. Section 6 conducts an extensive numerical calibration to select a smoothing parameter in the theoretically acceptable range, which results in a specific recommendation. Section 7 runs a full-blown Monte Carlo simulation exercise that demonstrates the strong performance of quadratic shrinkage in a wide variety of scenarios, matching the best RMT-based estimators that allow infinite degrees of nonlinearity. Section 8 concludes. The supplementary material Ledoit and Wolf (2022) contains programming code, proofs of all mathematical results, and further material such as additional Monte Carlo simulations.

2. Finite-sample analysis

Even though the sample size n is fixed in this section, we nonetheless subscript quantities by n to harmonize the notation throughout the paper.

2.1. General setup

Let Σ_n denote a p-dimensional positive-definite population covariance matrix, where $p < n$.[1] A mean-zero independent and identically distributed (i.i.d.) sample of n observations with covariance matrix Σ_n is arranged in an $n \times p$ matrix Y_n, which generates the sample covariance matrix $S_n := Y_n' Y_n / n$.[2] Its spectral decomposition is $S_n = U_n \Lambda_n U_n'$, where Λ_n is the diagonal matrix whose elements are the eigenvalues $\boldsymbol{\lambda}_n := (\lambda_{n,1}, \ldots, \lambda_{n,p})$ sorted in nondecreasing order without loss of generality, and U_n is an orthogonal matrix whose columns $[u_{n,1} \ldots u_{n,p}]$ are the corresponding eigenvectors. Thus, $S_n = \sum_{i=1}^{p} \lambda_{n,i} \cdot u_{n,i} u_{n,i}'$.

2.2. Class of estimators

Following Stein (1986, Lecture 4), we seek an estimator of the form $\widetilde{\Sigma}_n := U_n \widetilde{\Delta}_n U_n'$, where $\widetilde{\Delta}_n$ is a diagonal matrix whose elements $\widetilde{\boldsymbol{\delta}}_n := \left(\widetilde{\delta}_{n,1}, \ldots, \widetilde{\delta}_{n,p}\right) \in (0, +\infty)^p$ are a function of $\boldsymbol{\lambda}_n$. Such estimators are *rotation* equivariant because post-multiplying the data Y_n by an orthogonal matrix (with determinant one) rotates the estimators accordingly. By contrast, estimators from the sparsity literature such as the ones of Bickel and Levina (2008a,b) and El Karoui (2008) are merely *permutation* equivariant. This means that they are dependent on *a priori* information about the orientation of the orthonormal basis of the population eigenvectors, which is impossible to verify in practice.

2.3. Loss function

Any generic estimator $\widetilde{\Sigma}_n$ is evaluated according to the following loss function used by Stein (1975, 1986) and commonly referred to as Stein's loss:

Definition 2.1 (Stein's Loss). Let $\mathsf{Tr}(\cdot)$ denote the trace. Stein's loss is defined as:

$$\mathcal{L}_n^{\mathrm{ST}}\left(\Sigma_n, \widetilde{\Sigma}_n\right) := \frac{1}{p}\mathsf{Tr}\left(\Sigma_n^{-1}\widetilde{\Sigma}_n\right) - \frac{1}{p}\log\det\left(\Sigma_n^{-1}\widetilde{\Sigma}_n\right) - 1. \tag{2.1}$$

Proposition 2.1. *The solution to the optimization problem*

$$\underset{\widetilde{\Delta}_n \text{ diagonal}}{\arg\min} \quad \mathcal{L}_n^{ST}\left(\Sigma_n, U_n \widetilde{\Delta}_n U_n'\right)$$

$$\textit{is} \quad \overline{D}_n^{ST} := \mathsf{Diag}\left(\overline{d}_{n,1}^{ST}, \ldots, \overline{d}_{n,p}^{ST}\right) \quad \textit{where} \quad \overline{d}_{n,i}^{ST} := \frac{1}{u_{n,i}' \Sigma_n^{-1} u_{n,i}} \quad \textit{for } i = 1, \ldots, p. \tag{2.2}$$

This results in an estimator, $\overline{S}_n^{\mathrm{ST}} := U_n \overline{D}_n^{\mathrm{ST}} U_n'$, which is not achievable in practice because Σ_n is unobservable, but constitutes a useful benchmark. At the qualitative level, we can already point out that the shrinkage formula (2.2) is in some sense 'the inverse of an inverse'. A less roundabout, more direct approach will be provided in Section 4.

[1] The singular case $p > n$ is covered in Section 5 and in Section E of the supplementary material Ledoit and Wolf (2022).

[2] If the variables have a nonzero mean, the theorems in this paper also apply after in-sample demeaning, and adjusting the 'effective sample size' to $n - 1$ to account for the loss of one degree of freedom.

2.4. *Bona fide* shrinkage formula

Rewriting Equations (22)–(23) of Stein (1986, Lecture 4) in our notation, Stein approximates the unobservable $\overline{d}_{n,i}^{\mathrm{ST}}$'s with the *bona fide* estimator

$$\widetilde{d}_{n,i} := \frac{n\lambda_{n,i}}{n + p - 2i + 1 + 2\displaystyle\sum_{j=i+1}^{p} \frac{n\lambda_{n,j}}{n\lambda_{n,i} - n\lambda_{n,j}} - 2\sum_{j=1}^{i-1} \frac{n\lambda_{n,i}}{n\lambda_{n,j} - n\lambda_{n,i}}}, \tag{2.3}$$

for $i = 1, \dots, p$. The main differences between Stein's notation and ours are that the roles of the indices i and j are swapped, and his eigenvalues ℓ_j $(j = 1, \dots, p)$ are equal to n times those of the sample covariance matrix. The resulting covariance matrix estimator is $\widetilde{S}_n := \sum_{i=1}^{p} \widetilde{d}_{n,i} \cdot u_{n,i} u_{n,i}'$. Although expression (2.6) below is the more common one in the literature, the original expression in Stein (1986, Lecture 4) is indeed (2.3).

Stein's estimator broke new ground and fathered an extensive literature on rotation-equivariant shrinkage estimation of a covariance matrix; for example, see the articles by Chen, Wiesel and Hero (2009), Daniels and Kass (2001), Dey and Srinivasan (1985), Donoho, Gavish and Johnstone (2018), Efron and Morris (1976), Haff (1980, 1991), Krishnamoorthy and Gupta (1989), Ledoit and Wolf (2004, 2012), Lin and Perlman (1985), Loh (1991), Pal (1993), Won *et al.* (2013), Yang and Berger (1994), and the references therein.

2.5. The Stein shrinker

One glaring issue with Stein's formula is that it is not intuitive, being highly nonlinear. Therefore, our first contribution is to reinterpret it as linear shrinkage — but not of the sample eigenvalues, of their *inverses* instead. There is no *a priori* reason why linearly shrinking the inverse eigenvalues should be better than shrinking the eigenvalues themselves; it is just what Stein's mathematical discoveries implied. But linearly shrinking the inverse eigenvalues is certainly no more worrisome than linearly shrinking the eigenvalues, an operation that has been well understood and accepted by researchers at least since Ledoit and Wolf (2004).

Theorem 2.1. *The nonlinear shrinkage formula* (2.3) *is mathematically equivalent to*

$$\forall i = 1, \dots, p \qquad \widetilde{d}_{n,i}^{-1} = \left(1 - \frac{p-1}{n}\right) \lambda_{n,i}^{-1} + \left(\frac{p-1}{n}\right) \times 2\lambda_{n,i}^{-1} \widetilde{\theta}_n\left(\lambda_{n,i}^{-1}\right) \tag{2.4}$$

$$\textit{where} \quad \forall x \in \mathbb{R} \quad \widetilde{\theta}_n(x) := \frac{1}{p-1} \sum_{\substack{j=1 \\ \lambda_{n,j}^{-1} \neq x}}^{p} \lambda_{n,j}^{-1} \frac{1}{\lambda_{n,j}^{-1} - x}. \tag{2.5}$$

Proof. From Equation (2.3) we deduce

$$\widetilde{d}_{n,i} = \frac{\lambda_{n,i}}{1 + \dfrac{p-1}{n} + \dfrac{2}{n} \displaystyle\sum_{j \neq i} \frac{\lambda_{n,j}}{\lambda_{n,i} - \lambda_{n,j}}} \tag{2.6}$$

$$\widetilde{d}_{n,i}^{-1} = \lambda_{n,i}^{-1} + \frac{p-1}{n}\lambda_{n,i}^{-1}\left[1 + 2\frac{1}{p-1}\sum_{j\neq i}\left(\frac{\lambda_{n,j}^{-1}}{\lambda_{n,j}^{-1} - \lambda_{n,i}^{-1}} - 1\right)\right] \tag{2.7}$$

$$= \left(1 - \frac{p-1}{n}\right)\lambda_{n,i}^{-1} + \left(\frac{p-1}{n}\right) \times \frac{1}{p-1}\sum_{j\neq i}\frac{2\lambda_{n,j}^{-1}\lambda_{n,i}^{-1}}{\lambda_{n,j}^{-1} - \lambda_{n,i}^{-1}}. \tag{2.8}$$

□

We call this shrinkage "linear" in inverse-eigenvalues space because it is a convex linear combination of $\lambda_{n,i}^{-1}$ with a shrinkage target that has a common structure independent of the shrinkage intensity. The fact that the shrinkage intensity is $(p-1)/n$ makes intuitive sense because more shrinkage must be applied when the curse of dimensionality is stronger. The shrinkage target is a multiplicative modulation of the eigenvalue $\lambda_{n,i}^{-1}$ that is being shrunk. From now on we shall call the modulator

$$\widetilde{\theta}_n(x) = \frac{1}{p-1}\sum_{\substack{j=1 \\ \lambda_{n,j}^{-1}\neq x}}^{p} \lambda_{n,j}^{-1}\frac{1}{\lambda_{n,j}^{-1} - x} \tag{2.9}$$

the "Stein shrinker". It is locally adaptive in the sense that $\widetilde{\theta}_n(x)$ is not constant. This property stands in sharp contrast with the linear shrinkage formula of Ledoit and Wolf (2004) that shrinks all sample eigenvalues linearly towards the same common (global) target: their grand mean. Letting shrinkage targets adapt to local conditions makes it possible to extract additional accuracy gains over and above those already attained by Ledoit and Wolf (2004), especially when the sample eigenvalues are dispersed or clustered.

Visual inspection of the Stein shrinker immediately reveals that: (1) it attracts eigenvalues towards each other; (2) larger precision matrix eigenvalues have proportionally stronger power of attraction; (3) the intensity of the attraction vanishes to zero as the distance between eigenvalues increases; and (4) bad things happen (explosive numerical behavior) when two eigenvalues get too close to each other. The first three properties are features, but the fourth one can be considered a 'bug', and the next section is devoted to fixing it by smoothing.[3]

3. Optimal linear-inverse shrinkage

Stein's analysis was only suggestive of optimality, not conclusive. First, he conceded that he had to ignore the effect of a certain derivatives term (Stein, 1986, p. 1391). Second, he post-processed the shrunken eigenvalues of Equation (2.3) through a numerical procedure called "isotonization" in order to restore the ordering of the eigenvalues, and to ensure that they are all positive. Rajaratnam and Vincenzi (2016b) show that isotonization is actually essential to the empirical success of Stein's estimator, but its theoretical properties are extremely hard to investigate formally. Our next task is, therefore, to develop a *provably* optimal version of Stein's estimator that is purely analytical in nature. Given the lack of tractability in finite samples, we move to the framework of *large-dimensional asymptotics*.

[3] Rajaratnam and Vincenzi (2016a,b) provide an in-depth study of the limitations of the original (or raw) Stein shrinker.

3.1. Large-dimensional asymptotics

The idea is that the matrix dimension p and sample size n go to infinity together, while their ratio p/n (called the "concentration (ratio)") converges to some limit $c \in (0,1)$.[4] This framework is empirically relevant as soon as the matrix dimension is non-negligible with respect to the sample size. The following set of assumptions, or slight variations thereof, have been employed before in this literature also known, especially in physics, as random matrix theory (RMT), going back to Wigner (1955).

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

The elegant way to handle the ever-increasing dimension of the vector of eigenvalues is to map it into a function:

Definition 3.1. The empirical distribution function (e.d.f.) of a collection of eigenvalues $(\alpha_1,\dots,\alpha_p)$ is the nondecreasing step function $x \longmapsto p^{-1}\sum_{i=1}^{p} \mathbb{1}_{\{\alpha_i \le x\}}$, where $\mathbb{1}$ denotes the indicator function.

This e.d.f. returns the proportion of eigenvalues that lie weakly below its argument.

Assumption 3.2 (Data Generating Process).

a. The population covariance matrix Σ_n is a nonrandom symmetric positive-definite matrix of dimension $p \times p$.
b. X_n is an $n \times p$ matrix of i.i.d. random variables with mean zero, variance one, and finite 16th 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.
c. Let $\boldsymbol{\tau}_n := (\tau_{n,1},\dots,\tau_{n,p})'$ 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)".
d. $\mathsf{Supp}(H)$, the support of H, is the union of a finite number of closed intervals, bounded away from zero and infinity. Also, there exists a compact interval $[\underline{T},\overline{T}] \subset (0,\infty)$ that contains $\left\{\tau_{n,1},\dots,\tau_{n,p}\right\}$ for all n large enough.

Remark 3.1. The assumption of a finite 16th moment comes from Theorem 3 of Jing *et al.* (2010), which we use in our proofs. However, these authors' Remark 1 conjectures that a finite fourth moment is actually enough, and our own Monte Carlo simulations in Section 7 concur. □

The literature on the eigenvalues of the sample covariance matrix under large-dimensional asymptotics is based on a foundational result by Marčenko and Pastur (1967), which has been strengthened and broadened by subsequent authors including Silverstein and Bai (1995) and Silverstein (1995), among others. The latter's Theorem 1.1 implies that, under Assumptions 3.1–3.2, there exists a continuous non-stochastic limiting sample spectral distribution function F such that the e.d.f. of the sample eigenvalues, denoted by F_n, converges pointwise almost surely to F. This limiting sample spectral c.d.f. F is uniquely determined by c and H, so we will denote it more explicitly by $F_{c,H}$ whenever there is some

[4] The case $c \in (1,+\infty)$ is covered in Section 5 and in Section E of the supplementary material Ledoit and Wolf (2022).

risk of ambiguity. Assumption 3.2 together with Theorem 1.1. of Bai and Silverstein (1998) implies that the support of F, denoted by $\mathsf{Supp}(F)$, is the union of a finite number $\nu \geq 1$ of compact intervals: $\mathsf{Supp}(F) = \bigcup_{k=1}^{\nu} [a_k, b_k]$, where $0 < a_1 < b_1 < \cdots < a_\nu < b_\nu < \infty$.

Following Ledoit and Wolf (2018), we extend the class of rotation-equivariant covariance matrix estimators from Section 2 into the realm of large-dimensional asymptotics.

Definition 3.2 (Class of Estimators). Covariance matrix estimators are of the type $\widetilde{\Sigma}_n := U_n \widetilde{\Delta}_n U_n'$, where $\widetilde{\Delta}_n$ is a diagonal matrix: $\widetilde{\Delta}_n := \mathsf{Diag}\left(\widetilde{\delta}_n(\lambda_{n,1}) \ldots, \widetilde{\delta}_n(\lambda_{n,p})\right)$, and $\widetilde{\delta}_n$ is a (possibly random) real univariate function which may depend on S_n.

Every candidate shrinkage function $\widetilde{\delta}_n$ must behave well asymptotically:

Assumption 3.3 (Limiting Shrinkage Function). There exists a nonrandom real univariate function $\widetilde{\delta}$ defined on $\mathsf{Supp}(F)$ and continuously differentiable such that $\widetilde{\delta}_n(x) \xrightarrow{\text{a.s}} \widetilde{\delta}(x)$, for all $x \in \mathsf{Supp}(F)$. Furthermore, this convergence is uniform over $x \in \bigcup_{k=1}^{\nu} [a_k + \eta, b_k - \eta]$, 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 $x \in \bigcup_{k=1}^{\nu} [a_k - \eta, b_k + \eta]$, $\widetilde{\delta}_n(x)$ is uniformly bounded by $\widetilde{K}$ from above and by $1/\widetilde{K}$ from below, for n large enough.

3.2. Smoothed Stein shrinker

Our second contribution is to prove that a simple smoothing of the Stein shrinker yields an optimal estimator under large-dimensional asymptotics, even without requiring the variates to be normally distributed.

Theorem 3.1. *Suppose Assumptions 3.1–3.3 hold. Then, for any covariance matrix estimator $\widetilde{\Sigma}_n$ in the rotation-equivariant class of Definition 3.2, Stein's loss $\mathcal{L}_n^{ST}(\Sigma_n, \widetilde{\Sigma}_n)$ converges almost surely to a nonrandom limit as p and n go to infinity together. This limit is minimized if $\widetilde{\delta}_n(\lambda_{n,i}) = \widehat{d}_{n,i}$, with $\widehat{d}_{n,i}$ satisfying:*

$$\forall i = 1, \ldots, p \qquad \widehat{d}_{n,i}^{-1} := \left(1 - \frac{p}{n}\right) \lambda_{n,i}^{-1} + \left(\frac{p}{n}\right) \times 2\lambda_{n,i}^{-1} \widehat{\theta}_n\left(\lambda_{n,i}^{-1}\right) \tag{3.1}$$

$$\textit{where} \quad \forall x \in \mathbb{R} \qquad \widehat{\theta}_n(x) := \frac{1}{p} \sum_{j=1}^{p} \lambda_{n,j}^{-1} \frac{\lambda_{n,j}^{-1} - x}{\left(\lambda_{n,j}^{-1} - x\right)^2 + h_n^2 \lambda_{n,j}^{-2}}, \tag{3.2}$$

$$\textit{and a smoothing parameter} \quad h_n \sim K n^{-\alpha} \quad \textit{for some } K > 0 \textit{ and } \alpha \in (0, 2/5). \tag{3.3}$$

The resulting covariance matrix estimator is $\widehat{S}_n := \sum_{i=1}^{p} \widehat{d}_{n,i} \cdot u_{n,i} u_{n,i}'$.

From here on, all proofs are in the the supplementary material Ledoit and Wolf (2022). As a technical aside, the proofs in this paper build upon Jing *et al.* (2010) and Ledoit and Wolf (2020). We use the most salient elements of both works as stepping stones to make further headway. From Jing *et al.* (2010), we borrow techniques that enable us to extend the analysis of Ledoit and Wolf (2020) from kernels with bounded support to kernels with unbounded support. From Ledoit and Wolf (2020), we borrow techniques that enable us to extend the kernel estimation of the limiting sample spectral density in Jing *et al.* (2010) to its Hilbert transform. Beyond both papers, we move into the realm of the *inverses* of

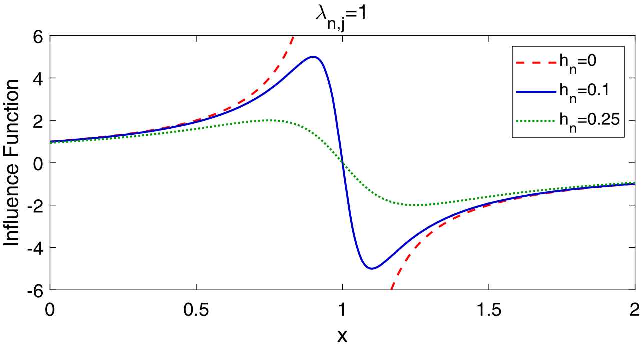

Figure 1. Dependence of the influence function on the regularization parameter h_n. When $h_n = 0$, the function diverges, which generates much numerical instability.

the sample eigenvalues (as opposed to the sample eigenvalues themselves), and of the first incomplete moment function.

Equation (3.1) is still linear shrinkage of the inverse sample eigenvalues. Replacing the shrinkage intensity $(p-1)/n$ from Equation (2.4) with p/n is immaterial, as we operate under large-dimensional asymptotics. What matters is that inside the summation, the discontinuous, explosive influence function

$$\frac{1}{\lambda_{n,j}^{-1}-x} \quad \text{is replaced by the smoother equivalent} \quad \frac{\lambda_{n,j}^{-1}-x}{\left(\lambda_{n,j}^{-1}-x\right)^2+h_n^2\lambda_{n,j}^{-2}}. \tag{3.4}$$

We view the covariance matrix estimator $\widehat{S}_n$ of Theorem 3.1 as linear shrinkage in inverse-eigenvalues space, or linear-inverse shrinkage (LIS) for short. It can also be interpreted as linear shrinkage of the precision matrix.

The bandwidth parameter h_n controls the degree of smoothing. If h_n were equal to zero, the two fractions in Equation (3.4) would be mathematically identical. For the purpose of the proofs, we require h_n to be strictly positive but to vanish asymptotically as n goes to infinity. In terms of nomenclature, we call $\widehat{\theta}_n(x)$ the "smoothed Stein shrinker". Figure 1 illustrates visually. Stein's formula required *ex post* numerical regularization through the *ad hoc* procedure of isotonization. We avoid this problem by bringing regularization analytically inside the formula through the introduction of the parameter h_n into the smoothed shrinkage modulator $\widehat{\theta}_n(x)$. Not only does Theorem 3.1 formally prove optimality, but it does so without requiring normality, as can be seen from Assumption 3.2(b).

Careful examination of the influence function shows a phenomenon of 'local shrinkage'. If $\lambda_{n,i}^{-1}$ is slightly below $\lambda_{n,j}^{-1}$ ($x < 1$ in Figure 1), then the influence exerted by $\lambda_{n,j}^{-1}$ onto $\lambda_{n,i}^{-1}$ is positive, meaning that $\widehat{d}_{n,i}^{-1}$ will tend to go up towards $\lambda_{n,j}^{-1}$ (everything else being equal). Similarly, if $\lambda_{n,i}^{-1}$ is slightly above $\lambda_{n,j}^{-1}$ ($x > 1$ in Figure 1), then the influence exerted by $\lambda_{n,j}^{-1}$ onto $\lambda_{n,i}^{-1}$ is negative, meaning that $\widehat{d}_{n,i}^{-1}$ will tend to go down, *also* towards $\lambda_{n,j}^{-1}$. Thus, there is shrinkage in the sense that inverse eigenvalues tend to be attracted towards one another. But, unlike the linear shrinkage formula of Ledoit and Wolf (2004), this shrinkage is 'local' because the influence exerted by distant eigenvalues vanishes quickly. This is why this particular form of linear shrinkage can generate substantial improvements when the population eigenvalues are dispersed, clustered, or otherwise unruly.

At this stage, there is nothing guaranteeing that the shrunken inverse eigenvalues will be all strictly positive. This was also the case in the original Stein estimator, which is part of the reason why he resorted to isotonization, thereby fixing the problem. But since we would prefer to retain a purely analytical formula, we propose a minor alteration to Equation (3.1). It is based on the observation that shrinkage always operates inwards at the extremities of the support. Thus, in particular, all $\widehat{d}_{n,i}^{-1}$ should be greater than or equal to $\min_{j=1,\dots,p}(\lambda_{n,j}^{-1}) = \lambda_{n,p}^{-1}$, which is itself strictly positive. Hence the correction

$$\left(\widehat{d}_{n,i}^{\mathrm{LIS}}\right)^{-1} := \max\left[\lambda_{n,p}^{-1}, \frac{1}{p}\sum_{j=1}^{p}\lambda_{n,j}^{-1}\frac{\lambda_{n,j}^{-1}-x}{\left(\lambda_{n,j}^{-1}-x\right)^2+h_n^2\lambda_{n,j}^{-2}}\right] \quad \text{and} \quad \widehat{S}_n^{\mathrm{LIS}} := \sum_{i=1}^{p}\widehat{d}_{n,i}^{\mathrm{LIS}}\cdot u_{n,i}u_{n,i}'. \qquad (3.5)$$

This modification is not needed asymptotically, so the constrained estimator shares the optimality properties stated in Theorem 3.1. Yet, we prefer it because it is safer in finite samples.[5]

4. Shrinkage under alternative loss functions

Stein (1986, p. 1390) candidly motivated the use of the loss function of Definition 2.1, which in the meantime has been called Stein's loss, "because it is comparatively easy to work with this loss function." The technological boost from large-dimensional asymptotics now gives us a chance to solve the same problem for loss functions that are "comparatively hard to work with", but potentially more interesting for a wide variety of practical applications.

4.1. Inverse Stein's loss

To facilitate continuity with Section 3, we start by applying Stein's loss to the precision matrix, as Tsukuma (2005) did.

Definition 4.1 (Inverse Stein's Loss). The Inverse Stein's loss is defined as:

$$\mathcal{L}_n^{IS}\left(\Sigma_n,\widetilde{\Sigma}_n\right) := \mathcal{L}_n^{ST}\left(\widetilde{\Sigma}_n,\Sigma_n\right) = \mathcal{L}_n^{ST}\left(\Sigma_n^{-1},\widetilde{\Sigma}_n^{-1}\right) = \frac{1}{p}\mathsf{Tr}\left(\Sigma_n\widetilde{\Sigma}_n^{-1}\right) - \frac{1}{p}\log\det\left(\Sigma_n\widetilde{\Sigma}_n^{-1}\right) - 1. \qquad (4.1)$$

Proposition 4.1. *The solution to the optimization problem*

$$\underset{\widetilde{\Delta}_n \text{ diagonal}}{\arg\min}\ \mathcal{L}_n^{IS}\left(\Sigma_n, U_n\widetilde{\Delta}_nU_n'\right)$$

$$\text{is}\quad \overline{D}_n := \mathsf{Diag}\left(\overline{d}_{n,1},\dots,\overline{d}_{n,p}\right) \quad \text{where} \quad \overline{d}_{n,i} := u_{n,i}'\Sigma_nu_{n,i} \quad \text{for } i=1,\dots,p. \qquad (4.2)$$

This results in an estimator, $\overline{S}_n := U_n\overline{D}_nU_n'$ which is not achievable in practice because Σ_n is unobservable, but constitutes a useful benchmark. For this reason, it is called an "oracle" estimator. Comparing with Stein's loss, in terms of covariance matrix eigenvalues we move from $\left(u_{n,i}'\Sigma_n^{-1}u_{n,i}\right)^{-1}$ to $u_{n,i}'\Sigma_nu_{n,i}$. This move appears rather intuitive and easily understandable.

[5] We thank two referees for having prompted us to find a way to guarantee the strict positivity of shrunken inverse eigenvalues.

4.2. Frobenius loss

Definition 4.2 (Frobenius Loss). The Frobenius loss is defined as:

$$\mathcal{L}^{\mathrm{FR}}(\Sigma_n, \widetilde{\Sigma}_n) := \frac{1}{p}\mathsf{Tr}\left[\left(\Sigma_n - \widetilde{\Sigma}_n\right)^2\right]. \tag{4.3}$$

The linear shrinkage formula of Ledoit and Wolf (2004) has popularized covariance matrix estimation based on the Frobenius loss in fields as far apart as cancer research (Pyeon *et al.*, 2007), macroeconomics (Korniotis, 2008), acoustics (Zhang, Sun and Zhang, 2009), brain-computer interface (Vidaurre *et al.*, 2009), psychology (Markon, 2010), political science (Tenenhaus and Tenenhaus, 2011), analytical chemistry (Guo *et al.*, 2012), geology (Elsheikh, Wheeler and Hoteit, 2013), and neuroscience (Deligianni *et al.*, 2014), to highlight but a selected few.

This literature also incorporates numerous extensions, adaptations and refinements of Ledoit and Wolf's linear shrinkage estimator. For example, Schäfer and Strimmer (2005) propose six different shrinkage targets; Stoica *et al.* (2008) embed linear shrinkage into space-time adaptive processing, an important radar technique; and Chen, Wiesel and Hero (2011) improve the formula for the shrinkage intensity when variates are normally, respectively elliptically, distributed. What such papers have in common is that they all set out to minimize the (expected) Frobenius loss.

Although not expected *a priori*, the solution for Frobenius loss turns out to be idential to the one for Inverse Stein's loss.

Proposition 4.2. *The oracle estimator* $\overline{S}_n := \sum_{i=1}^{p} \overline{d}_{n,i} \cdot u_{n,i}u_{n,i}'$ *with* $\overline{d}_{n,i}$ *defined as in* (4.2) *minimizes the loss function* $\mathcal{L}_n^{FR}$ *within the class of rotation-equivariant estimators specified in Section* 2.2.

As we will see now, there is yet another loss function leading to the same solution.

4.3. Minimum variance loss

An even stronger justification for moving away from Stein's loss comes from reviewing typical applications that make use of the covariance matrix and its inverse, in order to craft a tailor-made loss function that directly speaks to their overarching objectives. Markowitz (1952) essentially launched finance as a scientific field. His key contribution was to show that the investor should think in terms of a portfolio allocated across a multitude of candidate financial assets, and trade off expected returns (good) against total portfolio risk (bad). In this context, the variables whose covariances we estimate are asset returns (potentially in excess of the risk-free rate). Given a $p \times 1$ vector of expected (excess) returns μ, the optimal trade-off is achieved by solving the minimization problem

$$\underset{w\in\mathbb{R}^p}{\arg\min}\, w'\Sigma_n w \qquad \text{subject to: } w'\mu = \gamma, \tag{4.4}$$

where γ is an expected return target. This is also known as the "tangency portfolio". The solution to (4.4) is of the form $w = \text{scalar} \times \Sigma_n^{-1}\mu$, where the scalar multiplier is chosen to satisfy the investor's capital and leverage constraints. In reality, the population covariance matrix Σ_n is unobservable, so we must replace it with some estimator $\widetilde{\Sigma}_n$, and the question is how high is the *out-of-sample* volatility $\widetilde{w}'\Sigma_n\widetilde{w}$ of the *in-sample efficient* portfolio $\widetilde{w} := \widetilde{\Sigma}_n^{-1}\mu$. Given that every investor has a different μ based on individual expectations of how assets will perform, it is desirable to 'average out', or somehow 'integrate' the answer across all possible directions of the expected return vector μ.

The above framework carries through one-for-one as we move to Capon (1969) beamforming in signal processing (radar, sonar, wireless communications, seismology, etc). When a narrow-band source impinging upon an array of sensors, the vector μ should be interpreted as a response vector that includes effects such as coupling between elements and subsequent amplification, presumed to be known due to the design of the sensor array, called $\boldsymbol{a}(\boldsymbol{\theta}_s)$ by Abrahamsson, Selen and Stoica (2007) in their Equation (2). The rest of the analysis is identical to the finance application.

It is the same again for optimal fingerprinting, a method originally due to Hasselmann (1993) that has been chosen by the Intergovernment Panel on Climate Change to track global warming (IPCC, 2007, Section 9.A.1). In this context, the place of the vector μ is taken by the nonrandom response of the earth-system to an external forcing (Ribes, Azaïs and Planton, 2009, Section 2.1). The objective is to find a $p \times 1$ "fingerprint" vector so that the linear combination of temperature measurements weighted by the entries of the fingerprint minimizes climate-variability noise subject to a linear constraint on signal intensity.

Yet another method that fits in this framework is linear discriminant analysis (LDA), an essential tool for machine learning. In the two-class case, the vector μ represents the difference between the average score of one class across all dimensions of measurement, minus the average score of the other class against which we wish to discriminate. The objective of LDA is to find a one-dimensional subspace in which the classes are well separated. This is achieved by requiring that, after projection onto the subspace, the ratio of between-class variance to within-class variance is maximal. In this context, $\widetilde{w}$ is the direction of the one-dimensional subspace used to discriminate between classes. LDA has been used extensively in efforts to develop an interface between the brain and a computer (Vidaurre *et al.*, 2009).

Motivated by this seemingly ubiquitous mathematical problem, Engle, Ledoit and Wolf (2019) advocate what is called the "Minimum Variance" loss function:

Definition 4.3 (Minimum Variance Loss). The Minimum Variance loss is defined as:

$$\mathcal{L}^{\mathrm{MV}}\left(\Sigma_n, \widetilde{\Sigma}_n\right) := \frac{\mathsf{Tr}\left(\widetilde{\Sigma}_n^{-1}\Sigma_n\widetilde{\Sigma}_n^{-1}\right)/p}{\left[\mathsf{Tr}\left(\widetilde{\Sigma}_n^{-1}\right)/p\right]^2} - \frac{1}{\mathsf{Tr}\left[\Sigma_n^{-1}\right]/p}. \tag{4.5}$$

It represents the *true* variance of a linear combination of the original p variables selected to have minimum *estimated* variance subject to a generic linear constraint, suitably normalized under large-dimensional asymptotics. Thus, it is extremely relevant to the empirical applications listed above and mathematically similar ones in other fields of science. We will not go more in-depth here into the justification of this particular loss function because it has been given already in Engle, Ledoit and Wolf (2019, Section 4), as well as in a precursor paper by Engle and Colacito (2006, Section 2).

Proposition 4.3. *The oracle estimator* $\overline{S}_n := \sum_{i=1}^{p} \overline{d}_{n,i} \cdot u_{n,i}u_{n,i}'$ *with* $\overline{d}_{n,i}$ *defined as in* (4.2) *minimizes the loss function* $\mathcal{L}_n^{MV}$ *within the class of rotation-equivariant estimators specified in Section* 2.2.

The fact that the three loss functions $\mathcal{L}^{\mathrm{MV}}$, $\mathcal{L}^{\mathrm{FR}}$ and $\mathcal{L}_n^{IS}$, which have different motivations and different formulas, all lead to the same solution in terms of the oracle estimator serves as further justification for adopting this 'family'.

Remark 4.1. A common feature between all three loss functions in this family is that they are based on the population covariance matrix Σ_n itself, and not on its inverse Σ_n^{-1} used by Stein's loss. This feature is desirable because if one of the population eigenvalues happens to be close to zero, which can be hard to detect when $p > n$ as in Section 5, inverting Σ_n can generate much numerical instability. □

4.4. Conjugate and amplitude of the smoothed Stein shrinker

To construct a *bona fide* estimator, that is, one that only depends on the observable data Y_n, we go back to the more tractable framework of large-dimensional asymptotics. To express our solution, we need to present two closely related concepts borrowed from signal processing: the *conjugate* and the *amplitude*. As a sneak preview, the new loss functions will introduce a second shrinkage target governed by the squared amplitude of the smoothed Stein shrinker.

The notion of conjugate goes all the way back to the *analytic signal* theory of Gabor (1946). The basic idea is that what we observe (which, in our case, is the smoothed Stein shrinker $\widehat{\theta}_n(x)$) also encrypts a conjugate that is not directly observable but is extractable via the *Hilbert transform*. This important transform is defined as convolution with the Cauchy kernel $(\pi t)^{-1}$:

$$\forall x \in \mathbb{R} \qquad \widehat{\theta}_n^*(x) = \mathcal{H}_{\widehat{\theta}_n}(x) := \frac{1}{\pi} PV \int_{-\infty}^{+\infty} \widehat{\theta}_n(t) \frac{dt}{t-x}, \tag{4.6}$$

where PV stands for the *Cauchy principal value*, which is used to evaluate the singular integral:

$$PV \int_{-\infty}^{+\infty} \widehat{\theta}_n(t) \frac{dt}{t-x} := \lim_{\varepsilon \to 0^+} \left[\int_{-\infty}^{x-\varepsilon} \widehat{\theta}_n(t) \frac{dt}{t-x} + \int_{x+\varepsilon}^{+\infty} \widehat{\theta}_n(t) \frac{dt}{t-x} \right]. \tag{4.7}$$

Conjugation is anti-involutive, meaning that the conjugate of the conjugate is none other than the original function itself (up to a minus sign). For example, the conjugate of the sine is the cosine, and the conjugate of the cosine is minus the sine. So the two conjugates are best thought of as a pair 'joined at the hip'. In our case, it is worth finding out the interpretation of the conjugate of the Stein shrinker.

Proposition 4.4. *The smoothed Stein shrinker*

$$\widehat{\theta}_n(x) = \frac{1}{p} \sum_{j=1}^{p} \lambda_{n,j}^{-1} \frac{\lambda_{n,j}^{-1} - x}{\left(\lambda_{n,j}^{-1} - x\right)^2 + h_n^2 \lambda_{n,j}^{-2}} \tag{4.8}$$

has for its conjugate

$$\widehat{\theta}_n^*(x) := \frac{1}{p} \sum_{j=1}^{p} \lambda_{n,j}^{-1} \frac{h_n \lambda_{n,j}^{-1}}{\left(\lambda_{n,j}^{-1} - x\right)^2 + h_n^2 \lambda_{n,j}^{-2}}. \tag{4.9}$$

We see that (4.9) is a nonparametric estimator of the derivative of the first incomplete moment of the spectral distribution of the sample precision matrix, using the Cauchy density as kernel, and using the locally adaptive bandwidth $h_{n,j} := h_n \lambda_{n,j}^{-1}$. Thus, the conjugate $\widehat{\theta}_n^*(x)$ shows where the inverse eigenvalues lie, like a density that would overweight the larger inverse eigenvalues. It is intuitively satisfying that the attraction force between eigenvalues is the conjugate of their location, and vice-versa.

The analytic signal theory of Gabor (1946) goes one step further to define the *amplitude* of a signal by combining the original function with its conjugate in a quadratic way. Vakman (1996, Section II) proves Gabor's formula captures our physical intuition about what amplitude should mean. Formally, the squared amplitude is defined as follows:

$$\mathcal{A}_{\widehat{\theta}_n}^2(x) := \widehat{\theta}_n(x)^2 + \widehat{\theta}_n^*(x)^2 = \left[-\mathcal{H}_{\widehat{\theta}_n^*}(x) \right]^2 + \left[\mathcal{H}_{\widehat{\theta}_n}(x) \right]^2 \tag{4.10}$$

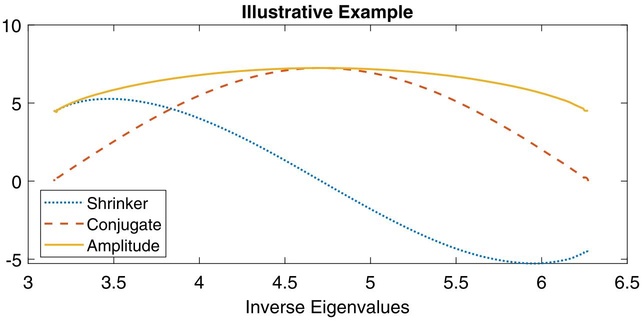

Figure 2. Stylized example of a shrinker, its conjugate and their amplitude.

$$= \left[\frac{1}{p} \sum_{j=1}^{p} \lambda_{n,j}^{-1} \frac{\lambda_{n,j}^{-1} - x}{\left(\lambda_{n,j}^{-1} - x\right)^2 + h_n^2 \lambda_{n,j}^{-2}} \right]^2 + \left[\frac{1}{p} \sum_{j=1}^{p} \lambda_{n,j}^{-1} \frac{h_n \lambda_{n,j}^{-1}}{\left(\lambda_{n,j}^{-1} - x\right)^2 + h_n^2 \lambda_{n,j}^{-2}} \right]^2 . \tag{4.11}$$

The amplitude measures whether there is any action in terms of either attraction force (the shrinker $\widehat{\theta}_n$) or weighted density (its conjugate $\widehat{\theta}_n^*$); it puts both pair members on the same footing and 'envelopes' them. Near the outskirts, there is less action of either type, so the amplitude vanishes. Note the elegant symmetry in (4.11): The denominator of the fraction is the sum of the squares of two terms, and both terms take turns appearing in the numerators of $\widehat{\theta}_n(x)$ and $\widehat{\theta}_n^*(x)$, respectively. To illustrate, Figure 2 displays a shrinker that behaves almost like minus the cosine over the interval $[\pi, 2\pi]$, up to rescaling. As expected, its conjugate is basically minus the sine (up to rescaling), with a peak at the support midpoint of $3\pi/2$. The amplitude looks relatively flat because the cosine (attraction force) and the sine (weighted density) tend to compensate for each other. Near the edges of the support, the amplitude starts dipping because the action is dying down. To recapitulate:

- In the middle of the cluster of inverse eigenvalues, the amplitude is strong because the weighted density (the conjugate) is strongly positive.
- On either side of the cluster of inverse eigenvalues, the amplitude is still strong, but for a different reason: because the attraction force (the shrinker) is strong in absolute value.
- Further out near the edges of the support, the inverse eigenvalues are sparser and exert less pull, so the amplitude weakens.

4.5. Quadratic-Inverse Shrinkage (QIS) estimator

We are now ready to state our third contribution, which is the adaptation of Stein's (smoothed) shrinkage formula to the Frobenius loss function and its two cousins:

Theorem 4.1. *Suppose Assumptions 3.1–3.3 hold. Then, for any covariance matrix estimator $\widehat{\Sigma}_n$ in the rotation-equivariant class of Definition 3.2, the Frobenius loss $\mathcal{L}_n^{FR}(\Sigma_n, \widehat{\Sigma}_n)$ converges in probability to a nonrandom limit as n goes to infinity. This limit is minimized if $\widetilde{\delta}_n(\lambda_{n,i}) = \widehat{\delta}_{n,i}$, with $\widehat{\delta}_{n,i}$ satisfying*

$$\widehat{\delta}_{n,i}^{-1} = \left(1 - \frac{p}{n}\right)^2 \lambda_{n,i}^{-1} + 2\frac{p}{n}\left(1 - \frac{p}{n}\right) \lambda_{n,i}^{-1} \widehat{\theta}_n\left(\lambda_{n,i}^{-1}\right) + \left(\frac{p}{n}\right)^2 \lambda_{n,i}^{-1} \mathcal{A}_{\widehat{\theta}_n}^2\left(\lambda_{n,i}^{-1}\right), \textit{ where} \tag{4.12}$$

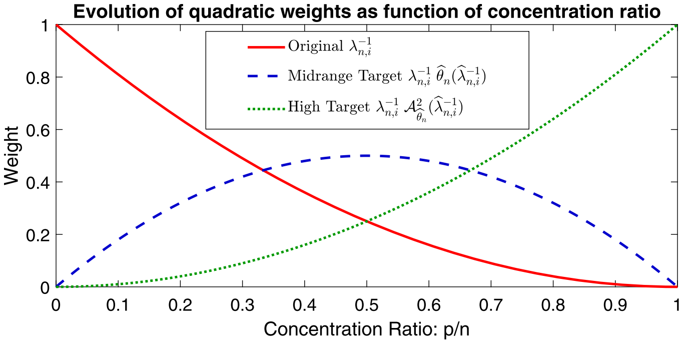

Figure 3. Evolution of the quadratic weights as a function of the concentration ratio p/n. The three of them sum up to one for every value of p/n.

$$\widehat{\theta}_n(x) := \frac{1}{p}\sum_{j=1}^{p}\lambda_{n,j}^{-1}\frac{\lambda_{n,j}^{-1}-x}{\left(\lambda_{n,j}^{-1}-x\right)^2+h_n^2\lambda_{n,j}^{-2}} \quad \textit{and} \tag{4.13}$$

$$\mathcal{A}^2_{\widehat{\theta}_n}(x) = \left[\frac{1}{p}\sum_{j=1}^{p}\lambda_{n,j}^{-1}\frac{\lambda_{n,j}^{-1}-x}{\left(\lambda_{n,j}^{-1}-x\right)^2+h_n^2\lambda_{n,j}^{-2}}\right]^2 + \left[\frac{1}{p}\sum_{j=1}^{p}\lambda_{n,j}^{-1}\frac{h_n\lambda_{n,j}^{-1}}{\left(\lambda_{n,j}^{-1}-x\right)^2+h_n^2\lambda_{n,j}^{-2}}\right]^2, \tag{4.14}$$

for a smoothing parameter h_n satisfying the conditions of Equation (3.3).

Note that the first two terms on the right-hand side of Equation (4.12) are the same as the ones from Theorem 3.1, up to rescaling by a factor of $1-(p/n)$. We call Equation (4.12) "quadratic shrinkage" of the inverse sample covariance matrix eigenvalues because the three weighting coefficients are quadratic functions of the concentration ratio p/n, adding up to a perfect square:

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

and also because the new term is the square of the amplitude of the smoothed Stein shrinker. Mathematically speaking, this new term goes back to the oracle formula for the Frobenius loss first proven in Ledoit and Péché (2011, Theorem 4); in particular, it is due to the squared modulus of the Stieltjes-transform term in their Equation (13).

Corollary 4.1. *The results stated in Theorem 4.1 also hold true for the Minimum Variance loss function $\mathcal{L}^{MV}$ and the Inverse Stein's loss function $\mathcal{L}^{IS}$, using the same shrinkage formula.*

The inverse sample covariance matrix eigenvalue is attracted to two shrinkage targets modulated, respectively, by $\widehat{\theta}_n(\lambda_{n,i}^{-1})$ and $\mathcal{A}^2_{\widehat{\theta}_n}(\lambda_{n,i}^{-1})$. The first target dominates mid-level concentration ratios and the second one high-level concentration ratios. Figure 3 illustrates these shapes.

Remark 4.2. The shrunken inverse eigenvalues $\{\widehat{\delta}_{n,i}^{-1}\}_{i=1}^{p}$ are guaranteed to be strictly positive, which was not necessarily the case when using Stein's loss in Section 2 and Theorem 3.1. Hence, a constraint analogous to Equation (3.5) is no longer required and shall not be applied here. As for undoing eigenvalue order violations through post-processing by a numerical algorithm, Section F.1 of the supplementary material Ledoit and Wolf (2022) shows that it is not particularly useful here because the violations are relatively few, their magnitudes are benign, and it would not really move the needle in terms of accuracy of the QIS estimator.

Theorem 4.1 states that the quadratic shrinkage formula is optimal among *all* nonlinear shrinkage formulas; therefore, it cannot be beaten by any cubic or higher-order shrinkage. This result that reduces the most complicated nonlinear problem to a simple quadratic by tweaking Stein's classic formula is both powerful and mathematically elegant.

A potentially valuable refinement, more useful for the Frobenius and Inverse Stein losses than for Minimum Variance loss, is to rescale the quadratic-inverse shrinkage estimator to have the same trace as the sample covariance matrix, a property already enjoyed by the linear shrinkage of Ledoit and Wolf (2004):

$$\widehat{\Sigma}_n^{\text{QIS}} := \frac{\mathsf{Tr}\big[S_n\big]}{\mathsf{Tr}\big[\widehat{\Sigma}_n\big]}\,\widehat{\Sigma}_n \qquad \text{where} \qquad \widehat{\Sigma}_n := \sum_{i=1}^{p} \widehat{\delta}_{n,i} \cdot u_{n,i} u_{n,i}' \,. \tag{4.15}$$

This modification is not needed asymptotically, but may boost finite-sample performance in some applications. The bandwidth parameter h_n will be further specified in the next section.

Remark 4.3. For researchers who prefer a loss function that considers eigenvalues close to zero as being 'as extreme' as eigenvalues close to infinity, the Symmetrized Kullback-Leibler divergence $\mathcal{L}_n^{\text{SKL}}\big(\Sigma_n, \widetilde{\Sigma}_n\big) := \frac{1}{2p}\mathsf{Tr}\big(\Sigma_n^{-1}\widetilde{\Sigma}_n + \Sigma_n\widetilde{\Sigma}_n^{-1}\big) - 1$ of Moakher and Batchelor (2006, Eq. (17.8)) is invariant to matrix inversion. It is a convenient alternative to the affine-equivariant geodesic norm on the manifold of positive-definite matrices and to the Log-Euclidian norm. The two latter norms are invariant to matrix inversion as well — see Förstner and Moonen (1999, Section 3.2) and Arsigny *et al.* (2006, p.413), respectively — but they are less tractable in the present framework. The rotation-equivariant covariance matrix estimator asymptotically optimal with respect to the Symmetrized Kullback-Leibler divergence is constructed by geometrically averaging Linear-Inverse shrinkage with Quadratic-Inverse shrinkage:

$$\widehat{S}^{\text{SKL}} := \sum_{i=1}^{p} \sqrt{\widehat{S}^{\text{LIS}} \times \widehat{\Sigma}^{\text{QIS}}} \cdot u_{n,i} u_{n,i}' \,. \tag{4.16}$$

This result is a direct consequence of Theorems 3.1 and 4.1, conducted in the spirit of Ledoit and Wolf (2018, Section 4.4). For reasons detailed in the next section, this approach only works for $p < n$, but not in the singular case $p > n$. □

Remark 4.4. It is worth noting that a loss function that pertains to the precision matrix such as Inverse Stein can have the same optimal shrinkage formula as a loss function that pertains to the covariance matrix such as Frobenius. Some loss functions, such as the three mentioned in the previous remark, are invariant to matrix inversion. In addition, it is possible to obtain an optimal shrinkage estimator of the covariance matrix with respect to Frobenius loss by shrinking the eigenvalues of the precision matrix. For all these reasons, one cannot ask whether we are shrinking the covariance matrix or the precision matrix because it is not a well-posed question. □

5. Singular case

We address the case $p > n$ by considering the inverses of the non-null sample eigenvalues only. The shrunken inverse eigenvalues are then given by:

$$\widehat{\delta}_{n,i}^{-1} := \begin{cases} \left(\dfrac{p}{n} - 1\right) \times \dfrac{1}{n} \displaystyle\sum_{j=p-n+1}^{p} \lambda_{n,j}^{-1} & i = 1, \ldots, p-n \\ \lambda_{n,i}^{-1} \mathcal{A}^2_{\widehat{\underline{\theta}}_n}(\lambda_{n,i}^{-1}) & i = p-n+1, \ldots, p \end{cases} \quad \text{where} \tag{5.1}$$

$$\widehat{\underline{\theta}}_n(x) := \frac{1}{n} \sum_{j=p-n+1}^{p} \lambda_{n,j}^{-1} \frac{\lambda_{n,j}^{-1} - x}{\left(\lambda_{n,j}^{-1} - x\right)^2 + h_n^2 \lambda_{n,j}^{-2}} \quad \text{and} \tag{5.2}$$

$$\mathcal{A}^2_{\widehat{\underline{\theta}}_n}(x) := \left[\frac{1}{n} \sum_{j=p-n+1}^{p} \lambda_{n,j}^{-1} \frac{\lambda_{n,j}^{-1} - x}{\left(\lambda_{n,j}^{-1} - x\right)^2 + h_n^2 \lambda_{n,j}^{-2}}\right]^2 + \left[\frac{1}{n} \sum_{j=p-n+1}^{p} \lambda_{n,j}^{-1} \frac{h_n \lambda_{n,j}^{-1}}{\left(\lambda_{n,j}^{-1} - x\right)^2 + h_n^2 \lambda_{n,j}^{-2}}\right]^2 . \tag{5.3}$$

Section E of the supplementary material Ledoit and Wolf (2022) goes through the derivations in more detail. This shrinkage formula also works for Minimum Variance and Inverse Stein loss. It should be post-processed through (4.15) as before.

Note that Equations (5.2)–(5.3) are the same as their counterparts in Theorem 4.1, with the proviso that averaging only extends over non-null eigenvalues. Also, if we compare Equation (4.12) with the bottom line of Equation (5.1), the first two terms on the right-hand side inherited from Stein's loss have disappeared and all the weight has shifted onto the third term. We consider this covariance matrix estimator to be the same as the one defined by Theorem 4.1, given that these two definitions cover two different domains: $p > n$ and $p < n$ respectively, so we can use the same symbol $\widehat{\Sigma}_n^{\text{QIS}}$ both times.

This approach cannot be used to estimate the covariance matrix under Stein's loss or the Symmetrized Kullback-Leibler divergence when $p > n$. The reason is that, as shown by Ledoit and Péché (2011, Theorem 5), adjusting null eigenvalues in these two cases requires estimating the population eigenvalues $(\tau_{n,1}, \ldots, \tau_{n,p})$, which is feasible only through numerical inversion of the QuEST function; see Ledoit and Wolf (2015, 2017). This feature constitutes yet another advantage of quadratic-inverse shrinkage.

There is a deep mathematical reason for the feature. If $p > n$, the number of null sample eigenvalues remains constant as long as the number of null population eigenvalues is less than $p - n$. Thus, the first $p - n$ null population eigenvalues are essentially undetectable. Only if there are $p - n + 1$ null population eigenvalues can we detect that there are indeed $p - n + 1$ null population eigenvalues, since we observe $p - n + 1$ null sample eigenvalues, and that is the only possible reason why. Thus, there is an initial zone where a few null population eigenvalues fly under the radar screen, followed by an 'avalanche effect' of revelation when all of a sudden there are too many of them. For this reason, any loss function that involves the inverse population covariance matrix should be avoided *at all cost* when $p > n$, and this includes both Stein's loss and the Symmetrized Kullback-Leibler divergence, as well as all the other loss functions that are invariant with respect to matrix inversion on the manifold of symmetric positive-definite matrices (tensors).

Remark 5.1. The case $p = n$ is not covered by the mathematical treatment for reasons similar to the ones outlined in Ledoit and Wolf (2015, Remark 3.2). But unreported Monte Carlo simulations show that, just like the QuEST method and the NERCOME method of Lam (2016), which are both of numerical nature, quadratic-inverse shrinkage also can successfully handle this case in practice. □

6. Smoothing parameter and concentration ratio

The only guidance from theory about selecting the smoothing parameter h_n is that it must be of the form $Kn^{-\alpha}$ for some constants $K > 0$ and $\alpha \in (0,2/5)$. However, this leaves open the question of dependency on $c_n := p/n$, the concentration ratio. Assumption 3.1 has it converging to some limit c bounded away from zero and infinity, so in theory the concentration is a constant that could appear anywhere in the formula. For example, we could have $h_n = K(c_n)\,n^{-\alpha}$, or even $h_n = K(c_n) \times [c_n n]^{-\alpha} = K(c_n)\,p^{-\alpha}$. In the absence of theoretical insight into this important question, we undertake a comprehensive numerical analysis to map out the terrain.

6.1. Test bench

We cover six different scenarios for the population spectrum (indexed $i = 1, \ldots, 6$ below):

1. Σ_n is the identity, and the variates are distributed as a "Student" t-distribution with 5 degrees of freedom.
2. High condition number: the population eigenvalues are uniformly distributed on the interval $[1,30]$; the variates are Gaussian.
3. Spectral separation: half of the population eigenvalues are uniformly distributed on the interval $[1,2]$, and the other half on $[10,11]$; the variates are still Gaussian.
4. Upper spike: the first $p-1$ population eigenvalues are distributed as per the semi-circular law (Wigner, 1955) with support $[1,5]$, and the top one is a "spike" that lies well above the bulk at the value 10; the variates are still Gaussian.
5. Lower spike: the top $p-1$ population eigenvalues are distributed as per the semi-circular law with support $[1,5]$, and the smallest one is a "reverse spike" that lies well below the bulk at the value 0.25; the variates are still Gaussian.
6. Skewed: many small eigenvalues, few large ones; based on the Marčenko and Pastur (1967) law with parameter 1/2, which implies a condition number ≈ 33; the variates are still Gaussian.

These six scenarios are furthermore crossed with two loss functions, Inverse Stein (Definition 4.1) and Minimum Variance (Definition 4.3), for a total of $6 \times 2 = 12$ combinations.

6.2. Monte Carlo simulation results

To vary the ratio p/n, we consider a collection of concentration ratios c defined as the tangents of angles in the set $\{10, 20, \ldots, 80\}$ (in degrees). To explore what happens for large and small sample sizes, we select 12 values for $\sqrt{pn}$ logarithmically spaced between 75 and 500. These two choices imply that p and n vary between a low of $75 \times \sqrt{\tan(10^\circ)} \approx 31$ and a high of $500 \times \sqrt{\tan(80^\circ)} \approx 1191$, a broad enough range. For each pair (p,n), we run 1,000 simulations.

For every (p,n) combination, for every population spectrum $i \in \{1, \ldots, 6\}$, and for every loss function $j \in \{\text{MV}, \text{IS}\}$, we determine numerically the optimal bandwidth $h_n^*(p,n,i,j)$. This is the value that minimizes the average of loss j across 1,000 Monte Carlo simulations run for the specification (p,n,i,j). Figure 4 displays the optimal bandwidth as a function of the matrix dimension and the sample size. The three axes are in logarithmic scale. Every point on the surface is the average of 12,000 numbers: 2 loss functions × 6 population spectra × 1,000 simulations. The optimal bandwidth decreases in p and/or n as expected, but its level, governed by the multiplier K, has a very clear dependence on the concentration ratio: it is high when p is close to n, and low otherwise. This makes sense because concentrations close

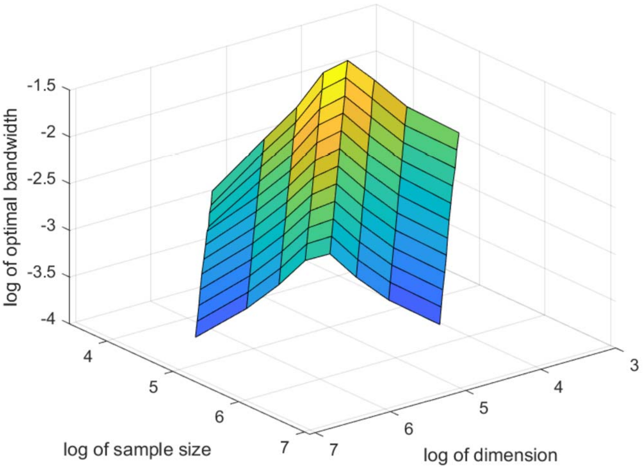

Figure 4. Dependence of the optimal bandwidth on the sample size and the matrix dimension.

to one generate many eigenvalues close to zero, which are hard to handle. This inverted V-shape pattern is universal: It looks exactly the same even if we partial out the results by loss function or by shape of the population spectrum. On the latter point, Section F.4 of the supplementary material Ledoit and Wolf (2022) provides further justification.

6.3. Numerical calibration of the optimal bandwidth

In order to formally model this structure, we run two regressions:

$$\log\left[h_n^*(p,n,i,j)\right] = a + b_1 \log\left[\min\left(\frac{p}{n},\frac{n}{p}\right)\right] + b_2 \log[n] + \varepsilon_{p,n,i,j} \tag{6.1}$$

$$\log\left[h_n^*(p,n,i,j)\right] = a + b_1 \log\left[\min\left(\frac{p}{n},\frac{n}{p}\right)\right] + b_2 \log[p] + \varepsilon_{p,n,i,j}. \tag{6.2}$$

Table 1 presents the results obtained with Matlab's `fitlm` routine.

One can clearly see that the matrix dimension provides a better fit than the sample size. Therefore, we are going into the direction of a bandwidth formula of the type $h_n = K(c_n)p^{-\alpha}$, rather than $h_n = K(c_n)n^{-\alpha}$. Looking at Model (6.2) specifically, the R^2 is very large at 86.4%, which justifies our modeling choice for the dependency of K on c_n, with a symmetric drop away from $p = n$ in the directions $p < n$ and $p > n$. One important aspect is that the exponent of p is close to the -0.40 boundary from theory, so we suggest rounding it to -0.35. Furthermore, the exponent of $\max(p/n, n/p)$ is quite close to twice this number: 0.70. Finally, the intercept is quite close to zero. Therefore, in the interest of elegance, our final recommendation for the bandwidth formula is simply:

$$\boxed{h_n := \underbrace{\min\left(\frac{p^2}{n^2},\frac{n^2}{p^2}\right)^{0.35}}_{K(c_n)} \times p^{-0.35}\ .} \tag{6.3}$$

Table 1. Fitting linear models (6.1) and (6.2) to the optimal bandwidth in log-space.

Model	(6.1)	(6.2)
Intercept	−1.021 (0.071)	0.256 (0.040)
$\log\left[\min(p/n, n/p)\right]$	0.659 (0.018)	0.659 (0.010)
$\log[n]$	−0.149 (0.013)	
$\log[p]$		−0.392 (0.007)
n	1152	1152
R^2	0.572	0.864

Based on more than a million Monte Carlo simulations, we consider this formula to be a safe all-around bandwidth that espouses the salient features of the problem at hand. Section F.4 of the supplementary material Ledoit and Wolf (2022) gives evidence that our proposed estimator is not sensitive to the choice of the smoothing hyper-parameter, as long as it remains in the general ballpark of Equation (6.3).

7. Numerical performance of quadratic-inverse shrinkage

Even though the main purpose of the present paper is essentially to exploit an unexpected connection between Stein's (1975) first-generation nonlinear shrinkage estimator of the covariance matrix and the latest advances in second-generation large-dimensional asymptotics, we still need to show how our proposed estimator — Quadratic-Inverse Shrinkage with smoothing parameter chosen as per Section 6 — performs relative to the current state-of-the-art. We put together a stable of seven competitors that do not assume any *a priori* information on the orientation of the eigenvectors of the true (unobservable) covariance matrix.

Sample The sample covariance matrix S_n.
Linear The linear shrinkage estimator of Ledoit and Wolf (2004).
NERCOME The nonlinear shrinkage estimator using cross-validation due to Lam (2016).
QIS The quadratic-inverse shrinkage estimator of Theorem 4.1 with smoothing parameter h_n chosen as per prescription (6.3).
Analytical The analytical nonlinear shrinkage formula of Ledoit and Wolf (2020, Section 4.7).
QuEST The nonlinear shrinkage estimator of Ledoit and Wolf (2015), which is based on numerical inversion of the QuEST function.
FSOPT The finite-sample optimal estimator $\overline{S}_n$ defined underneath Proposition 4.2, which would require knowledge of the population covariance matrix Σ_n, and thus is not applicable in the real world.

The first and the last estimator are used for benchmarking purposes, as they generate the percentage relative improvement in average loss (PRIAL), defined for any estimator $\widehat{\Sigma}_n$ as

$$\mathrm{PRIAL}_n\big(\widehat{\Sigma}_n\big) := \frac{\mathbb{E}\big[\mathcal{L}_n^{\mathrm{FR}}(S_n,\Sigma_n)\big] - \mathbb{E}\big[\mathcal{L}_n^{\mathrm{FR}}\big(\widehat{\Sigma}_n,\Sigma_n\big)\big]}{\mathbb{E}\big[\mathcal{L}_n^{\mathrm{FR}}(S_n,\Sigma_n)\big] - \mathbb{E}\big[\mathcal{L}_n^{\mathrm{FR}}\big(\overline{S}_n,\Sigma_n\big)\big]} \times 100\% \ . \tag{7.1}$$

Table 2. Simulation results for the baseline scenario.

Estimator	Sample	Linear	NERCOME	QIS	Analytical	QuEST	FSOPT
Average Loss	39.1	23.5	17.1	16.8	16.6	16.2	16.1
PRIAL	0%	68%	96%	97%	98%	99%	100%
Time (ms)	<1	1	2,117	3	3	1,644	3

The expectation $\mathbb{E}[\cdot]$ is in practice taken as the average across $\max\{100, \min\{1000, 10^5/p\}\}$ Monte Carlo simulations; for example, in dimension $p = 500$, we only need to run 200 simulations instead of 1000 to get reliable results.

From extant literature, we can already list some basic facts about these competitors.

- Linear shrinkage always beats the sample covariance matrix but, depending on the parameter configuration, can either get close to FSOPT or leave money on the table.
- The three nonlinear shrinkage estimators (NERCOME, Analytical, and QuEST) remain close to FSOPT in any parameter configuration, so it is very hard to beat them in terms of accuracy.
- Sample, Linear, Analytical, and FSOPT have closed-form expressions, so it is very hard to beat them in terms of speed and scalability in ultra-high dimensions; however, the numerical estimators NERCOME and QuEST are orders of magnitude slower.

Therefore, we will be able to qualify the QIS estimator as ‘state of the art’ if it has similar accuracy to NERCOME, Analytical, and QuEST; and similar speed/scalability to Sample, Linear, Analytical, and FSOPT. Also worth pointing out is that two of these estimators, namely Analytical and QuEST, are ‘outliers’ in the present context because their formulas derive not from statistics but from random matrix theory.

Even though the Monte Carlo simulations in this section focus on the Frobenius loss, because it has been widely accepted across many applied fields over the past couple of decades, additional simulations in Section F.3 of the supplementary material Ledoit and Wolf (2022) indicate that similar conclusions would carry over to the Inverse Stein’s loss and the Minimum Variance loss.

7.1. Baseline scenario

The simulations are organized around a baseline scenario. Each parameter will be subsequently varied to assess the robustness of the conclusions. The baseline scenario is:

- the matrix dimension is $p = 200$;
- the sample size is $n = 600$; therefore, the concentration ratio p/n is equal to $1/3$;
- the condition number of the population covariance matrix is 10;
- 20% of population eigenvalues are equal to 1, 40% are equal to 3, and 40% are equal to 10;
- and the variates are normally distributed.

The distribution of the population eigenvalues is a particularly interesting and difficult case introduced and analyzed in detail by Bai and Silverstein (1998). We have purposefully selected a shape of population spectrum left untouched by the calibration round in Section 6.

Table 2 presents estimator performances under the baseline scenario. Computational times come from a 3.3GHz Mac Pro desktop computer running Matlab R2020a.

The 0% PRIAL for the sample covariance matrix and the 100% PRIAL for the finite-sample optimal estimator are by construction. Linear shrinkage performs well but leaves some money on the

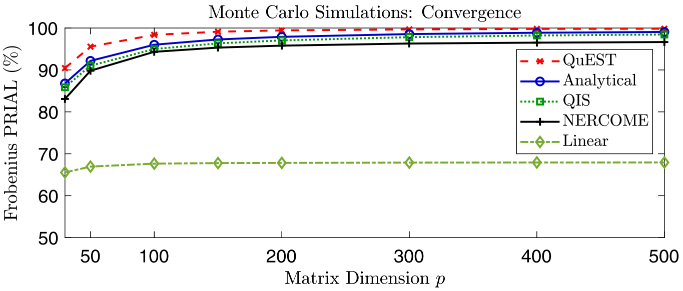

Figure 5. Evolution of the PRIAL as the matrix dimension and the sample size go to infinity together.

Table 3. Results of 100 Monte Carlo simulations for p = 10,000 and n = 30,000.

Estimator	Sample	Linear	QIS	Analytical	FSOPT
Average Loss	38.89	23.37	16.12	16.08	16.07
PRIAL	0%	68.0%	99.8%	99.9%	100%
Time (s)	5	10	31	32	35

table. The four nonlinear shrinkage formulas deliver near-perfect performance in the 96%+ range, with NERCOME and QuEST being much slower by orders of magnitude, as expected.

7.2. Convergence

Under large-dimensional asymptotics, the matrix dimension p and the sample size n go to infinity together, while their ratio p/n converges to some limit c. In the first experiment, p and n increase together, with their ratio fixed at the baseline value of $1/3$. Figure 5 displays the results. The four nonlinear shrinkage methods perform approximately the same as one another. They do well even in small dimensions, but do better as the dimension grows large. The difference between the PRIALs of Analytical and QIS is never more than 1%, which is very small.

Remark 7.1. Analytical, QIS and QuEST have shrinkage functions that behave essentially the same under large-dimensional asymptotics. They have the same limiting loss. Therefore, we should expect their performances to be nearly identical for large (p,n). For small and moderate (p,n), we would expect the performance of QuEST to be somewhat better compared to both Analytical and QIS, since it exploits the feature that f is the density of a limiting sample spectral distribution that is the output of the Fundamental Equation of Random Matrix Theory; hence, QuEST can be considered a model-based estimator. By contrast, Analytical and QIS do not exploit this feature of f, and thus can be considered model-free estimators. □

To see what happens when matrices become very large, we consider the case p = 10,000 in Table 3. At this level, the numerical methods QuEST and NERCOME can no longer follow, even with a powerful computer, so we only consider the other estimators.

One can see that letting the dimension go to infinity does nothing for linear shrinkage, but it brings nonlinear shrinkage ever closer to the maximum possible level of improvement, 100%. QIS is slightly

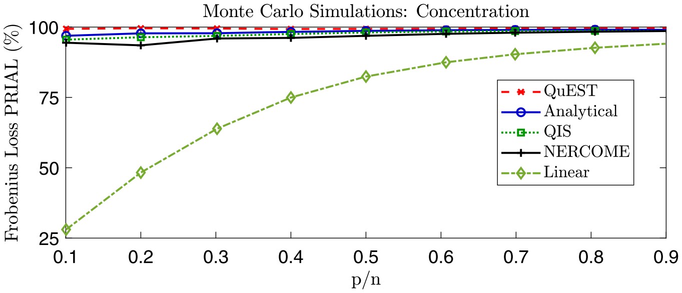

Figure 6. Evolution of the PRIAL of various estimators as a function of the ratio of the matrix dimension to the sample size.

worse than Analytical, but the difference is too small to be material. In unreported simulations, we even managed to go up to $p = 20{,}000$ and $n = 60{,}000$. The average computational time increased to just under four minutes per estimator, and was mostly devoted to extracting the sample covariance eigenvalues and eigenvectors.

7.3. Concentration ratio

We vary the concentration ratio p/n from 0.1 to 0.9 while holding the product $p \times n$ constant at the level it had under the baseline scenario, namely, $p \times n = 120{,}000$. (In this way, we keep the amount of total information, as measured by the number of entries in the matrix Y_n of Assumption 3.2b, fixed when the concentration ratio varies.) Figure 6 displays the resulting PRIALs. Linear shrinkage performs very well in high concentrations but less so in low concentrations, even though it still improves decisively over the sample covariance matrix across the board, as evidenced by its strictly positive PRIALs. The four nonlinear shrinkage methods perform approximately the same as one another, with QuEST remaining the gold standard, and the others performing nearly as well (for any practical purposes).

7.4. Condition number

We start again from the baseline scenario and, this time, vary the condition number of the population covariance matrix, called κ. We set 20% of the population eigenvalues equal to 1, 40% equal to $(2\kappa + 7)/9$, and 40% equal to κ. Thus, the baseline scenario corresponds to $\kappa = 10$. In this experiment, we let κ vary from $\kappa = 3$ to $\kappa = 30$. Doing so corresponds to linearly squeezing or stretching the distribution of population eigenvalues. Figure 7 displays the resulting PRIALs. Linear shrinkage performs very well for low condition numbers, but leaves some money on the table when eigenvalues are dispersed, as predicted theoretically by Ledoit and Wolf (2004, Figure 5). The nonlinear shrinkage formulas capture nearly all the potential for loss reduction.

7.5. Non-normality

In this experiment, we start from the baseline scenario and change the distribution of the variates. We study the Bernoulli coin toss distribution, which is the most platykurtic of all distributions, the Laplace distribution, which is leptokurtotic, and the Student t-distribution with 5 degrees of freedom, also

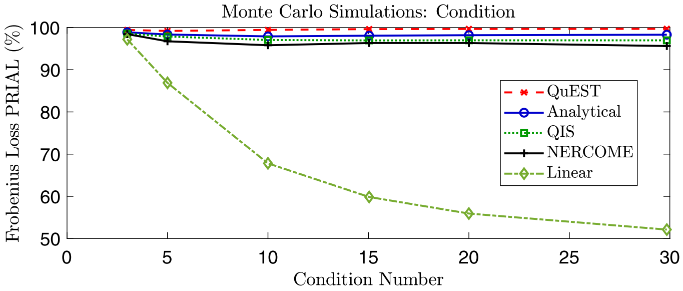

Figure 7. Evolution of the PRIAL of various estimators as a function of the condition number of the population covariance matrix.

Table 4. Simulation results for various variate distributions (PRIAL).

Distribution	Linear	NERCOME	QIS	Analytical	QuEST
Bernoulli	67.5%	95.9%	97.5%	98.1%	99.4 %
Laplace	68.4%	95.7%	96.4%	97.5%	99.1%
Student t_5	68.7%	95.7%	95.4%	96.5%	98.2 %

Table 5. Simulation results for various distributions of the population eigenvalues (PRIAL).

Beta Parameters	Linear	NERCOME	QIS	Analytical	QuEST
(1 , 1)	92.8 %	98.3 %	98.5 %	98.8 %	99.2 %
(1 , 2)	96.6 %	97.6 %	98.4 %	98.6 %	98.8 %
(2 , 1)	97.2 %	99.2 %	98.9 %	99.2 %	99.6 %
(1.5 , 1.5)	96.1 %	98.6 %	98.7 %	99.0 %	99.3 %
(0.5 , 0.5)	82.8 %	97.8 %	98.2 %	98.5 %	99.1 %
(5 , 5)	99.1 %	99.5 %	99.0 %	99.3 %	99.7 %
(5 , 2)	99.0 %	99.6 %	99.0 %	99.4 %	99.8 %
(2 , 5)	98.4 %	98.6 %	98.7 %	99.0 %	99.2 %

leptokurtotic. All of these are suitably normalized to have mean zero and variance one, if necessary. Table 4 presents the results. This experiment confirms that results are not sensitive to the distribution of the variates.

7.6. Shape of the distribution of population eigenvalues

Relative to the baseline scenario, we now move away from the clustered distribution for the population eigenvalues and instead employ continuous distributions from the Beta family. They are linearly shifted and stretched so that the support is $[1, 10]$. A graphical illustration of the densities of the various Beta shapes is in Ledoit and Wolf (2018, Figure 8.4). Table 5 presents the results. This time, linear shrinkage does much better overall, except perhaps for the bimodal shape (0.5,0.5). This is due to the fact that, in the other cases, the optimal nonlinear shrinkage formula happens to be almost linear. Nonlinear

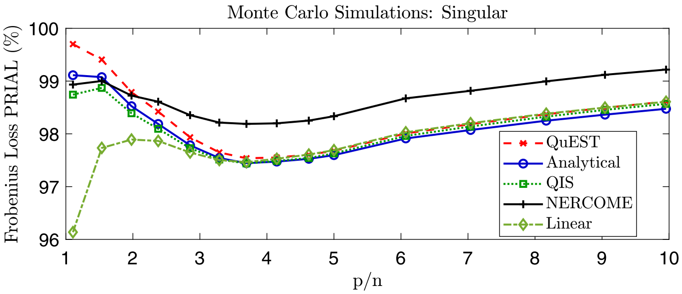

Figure 8. Frobenius PRIAL when the matrix dimension exceeds the sample size.

shrinkage formulas capture a nearly perfect percentage of the potential for variance reduction in all cases.

7.7. Singular case

Finally, we run a counterpart of the Monte Carlo simulations in Section 7.3 for the case $c > 1$. We vary the concentration ratio p/n from 1.1 to 10 while holding the product $p \times n$ constant at the level it had under the baseline scenario, namely, $p \times n = 120{,}000$. (In this way, we keep the amount of total information, as measured by the number of entries in the matrix Y_n of Assumption 3.2b, fixed when the concentration ratio varies.) Figure 8 displays the resulting PRIALs. We draw the attention of the reader to the vertical scale of the figure: It starts at 96%. This confirms the trend that could be inferred from Figure 6: Higher concentration ratios make all shrinkage estimators look good. At this level of performance, the exact ordering becomes relatively less important; overall, NERCOME does best.

7.8. Supplementary Monte Carlo simulations

To save space, some additional batches of simulations are relegated to Section F of the supplementary material Ledoit and Wolf (2022). Their conclusions are briefly summarized below for convenience:

- There is no real need to post-process the QIS estimator with a numerical algorithm that would restore the order of the shrunken eigenvalues.
- Choosing a smoothing kernel different from the Cauchy density would only increase complexity without increasing performance.
- The general pattern of simulation results presented in Section 7 for the Frobenius loss is similar for the Inverse Stein's loss and the Minimum Variance loss.
- The performance of the QIS estimator is not sensitive to the specification of the smoothing parameter h_n, as long as h_n remains in the general area of our proposal from Section 6.
- The classic estimator of Stein (1977, 1975, 1986) is still hard to beat on its own terms, that is, with respect to Stein's loss. The value of the QIS estimator is that it extends the same logic to alternative loss functions that are harder to handle mathematically, and potentially more attractive in practice (such as Frobenius loss and Minimum Variance loss), as well as being able to handle the challenging $p > n$ case.

7.9. Overall comparison of performance results

In terms of accuracy, the QIS estimator matches the high-water mark set by NERCOME, Analytical, and QuEST in hugging close to the FSOPT no matter what happens — unlike linear shrinkage, whose percentage improvement (albeit always positive) can fluctuate according to parameter configurations. In terms of speed and of scalability into ultra-high dimensions, the QIS estimator is in the efficient group alongside Sample, Linear, Analytical and FSOPT — a decisive advantage over the numerical methods NERCOME and QuEST.

If we define "state-of-the-art" as (i) close-to-FSOPT accuracy across the parameter space together with (ii) a scalable closed-formed mathematical expression, then QIS is currently the only estimator in this class apart from Analytical. (More such estimators, conceivably, could be invented by future researchers over the course of scientific progress.) What makes QIS, arguably, more attractive is that its formula

- originates from statistical decision theory (Stein, 1975), whereas the analytical shrinkage formula originates in the physics of random matrix theory (Marčenko and Pastur, 1967);
- is more intelligible because it is a simple adaptation of the Stein shrinker, which visibly attracts sample eigenvalues to close neighbors on either side, decaying with distance;
- and has a lower degree of complexity because it is second-order (quadratic) shrinkage, as opposed to infinite-order nonlinear shrinkage.

8. Conclusion

Stein's (1975,1977,1986) seminal work has garnered a lot of attention over the years from researchers interested in estimating covariance matrices of dimension larger than three. It is hard to make an original contribution on top of such a body of knowledge, but we (i) reinterpret Stein's ostensibly nonlinear shrinkage formula as linear in inverse-eigenvalues space; (ii) smooth out his shrinker to make it continuous instead of divergent; and (iii) address more practically-oriented loss functions that work even when variables outnumber observations by adjoining a quadratic component.

Given that this construct harnesses the latest techniques in large-dimensional asymptotic theory, we believe that it is not just tying up loose ends from the past, but also laying the foundations for a new covariance matrix estimator that will prove useful to future researchers. The relentless search for simpler formulas is a priority for the community because future developments will be easier to build on top of transparent insights instead of arcane ones. The intimate connection that we have established between one nonlinear shrinkage formula from the first generation (finite samples) and another nonlinear shrinkage formula from the second generation (large-dimensional asymptotics) is quite unexpected because of the wide gap in techniques and methodology; so the most likely reason is that a deeper mathematical truth has, at least partially, been unearthed.

Acknowledgements

We thank an associate editor and three anonymous referees for helpful comments that have greatly enhanced the exposition of the paper. We also thank Patrick Ledoit for acting in the capacity of a research assistant by writing the Python version of our Matlab programming code in Section A of the supplementary material Ledoit and Wolf (2022).

Supplementary Material

Programming code, proofs, and further material (DOI: 10.3150/20-BEJ1315SUPP; .pdf). This supplement contains programming code, proofs of all mathematical results, and further material such as additional Monte Carlo simulations.

References

Abrahamsson, R., Selen, Y. and Stoica, P. (2007). Enhanced covariance matrix estimators in adaptive beamforming. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing – ICASSP 2007* **II** 969–972.

Arsigny, V., Fillard, P., Pennec, X. and Ayache, N. (2006). Log-Euclidean metrics for fast and simple calculus on diffusion tensors. *Magn. Reson. Med.* **56** 411–421.

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 https://doi.org/10.1214/aop/1022855421

Bickel, P.J. and Levina, E. (2008a). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008 https://doi.org/10.1214/08-AOS600

Bickel, P.J. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969 https://doi.org/10.1214/009053607000000758

Bodnar, T., Gupta, A.K. and Parolya, N. (2016). Direct shrinkage estimation of large dimensional precision matrix. *J. Multivariate Anal.* **146** 223–236. MR3477661 https://doi.org/10.1016/j.jmva.2015.09.010

Capon, J. (1969). High-resolution frequency-wavenumber spectrum analysis. *Proc. IEEE* **57** 1408–1418.

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

Chen, Y., Wiesel, A. and Hero, A.O. III (2011). Robust shrinkage estimation of high-dimensional covariance matrices. *IEEE Trans. Signal Process.* **59** 4097–4107. MR2865971 https://doi.org/10.1109/TSP.2011.2138698

Chen, S.X., Zhang, L.-X. and Zhong, P.-S. (2010). Tests for high-dimensional covariance matrices. *J. Amer. Statist. Assoc.* **105** 810–819. MR2724863 https://doi.org/10.1198/jasa.2010.tm09560

Daniels, M.J. and Kass, R.E. (2001). Shrinkage estimators for covariance matrices. *Biometrics* **57** 1173–1184. MR1950425 https://doi.org/10.1111/j.0006-341X.2001.01173.x

Deligianni, F., Centeno, M., Carmichael, D.W. and Clayden, J.D. (2014). Relating resting-state fMRI and EEG whole-brain connectomes across frequency bands. *Front. Neurosci.* **8** 258. https://doi.org/10.3389/fnins.2014.00258

Dey, D.K. and Srinivasan, C. (1985). Estimation of a covariance matrix under Stein's loss. *Ann. Statist.* **13** 1581–1591. MR0811511 https://doi.org/10.1214/aos/1176349756

Donoho, D., Gavish, M. and Johnstone, I. (2018). Optimal shrinkage of eigenvalues in the spiked covariance model. *Ann. Statist.* **46** 1742–1778. MR3819116 https://doi.org/10.1214/17-AOS1601

Efron, B. and Morris, C. (1976). Multivariate empirical Bayes and estimation of covariance matrices. *Ann. Statist.* **4** 22–32. MR0394960

El Karoui, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.* **36** 2717–2756. MR2485011 https://doi.org/10.1214/07-AOS559

Elsheikh, A.H., Wheeler, M.F. and Hoteit, I. (2013). An iterative stochastic ensemble method for parameter estimation of subsurface flow models. *J. Comput. Phys.* **242** 696–714. MR3062055 https://doi.org/10.1016/j.jcp.2013.01.047

Engle, R. and Colacito, R. (2006). Testing and valuing dynamic correlations for asset allocation. *J. Bus. Econom. Statist.* **24** 238–253. MR2234449 https://doi.org/10.1198/073500106000000017

Engle, R.F., Ledoit, O. and Wolf, M. (2019). Large dynamic covariance matrices. *J. Bus. Econom. Statist.* **37** 363–375. MR3948411 https://doi.org/10.1080/07350015.2017.1345683

Förstner, W. and Moonen, B. (1999). A metric for covariance matrices. In *Quo Vadis Geodesia . . . ? Festschrift for Erik W. Grafarend on the Occasion of His 60th Birthday* (F. Krumm and V.S. Schwarze, eds.). *Technical Reports of the Department of Geodesy and Geoinformatics* **1999.6** 113–128. Univ. Stuttgart, Institute of Geodesy.

Gabor, D. (1946). Theory of communication. Part 1: The analysis of information. *J. Inst. Electr. Eng.* **93** 429–441.

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

Haff, L.R. (1979). Estimation of the inverse covariance matrix: Random mixtures of the inverse Wishart matrix and the identity. *Ann. Statist.* **7** 1264–1276. MR0550149

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

Haff, L.R. (1991). The variational form of certain Bayes estimators. *Ann. Statist.* **19** 1163–1190. MR1126320 https://doi.org/10.1214/aos/1176348244

Hasselmann, K. (1993). Optimal fingerprints for the detection of time-dependent climate change. *J. Climate* **6** 1957–1971.

Stein, C. (1977). Lectures on the theory of estimation of many parameters (in Russian). In *Studies in the Statistical Theory of Estimation, Part I* (I.A. Ibragimov and M.S. Nikulin, eds.). *Proceedings of Scientific Seminars of the Steklov Institute, Leningrad Division* **74** 4–65.

IPCC (2007). Climate change 2007: The physical science basis. In *Working Group I Contribution to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (S. Solomon, D. Qin, M. Manning, M. Marquis, K. Averyt, M.M.B. Tignor, H.L. Miller and Z. Chen, eds.) **4**. Cambridge and New York: Cambridge Univ. Press. p. 996.

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

Jing, B.-Y., Pan, G., Shao, Q.-M. and Zhou, W. (2010). Nonparametric estimate of spectral density functions of sample covariance matrices: A first step. *Ann. Statist.* **38** 3724–3750. MR2766866 https://doi.org/10.1214/10-AOS833

John, S. (1971). Some optimal multivariate tests. *Biometrika* **58** 123–127. MR0275568 https://doi.org/10.1093/biomet/58.1.123

Korniotis, G.M. (2008). Habit formation, incomplete markets, and the significance of regional risk for expected returns. *Rev. Financ. Stud.* **21** 2139–2172.

Krishnamoorthy, K. and Gupta, A.K. (1989). Improved minimax estimation of a normal precision matrix. *Canad. J. Statist.* **17** 91–102. MR1014094 https://doi.org/10.2307/3314766

Lam, C. (2016). Nonparametric eigenvalue-regularized precision or covariance matrix estimator. *Ann. Statist.* **44** 928–953. MR3485949 https://doi.org/10.1214/15-AOS1393

Ledoit, O. and Péché, S. (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probab. Theory Related Fields* **151** 233–264. MR2834718 https://doi.org/10.1007/s00440-010-0298-3

Ledoit, O. and Wolf, M. (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Ann. Statist.* **30** 1081–1102. MR1926169 https://doi.org/10.1214/aos/1031689018

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.* **88** 365–411. MR2026339 https://doi.org/10.1016/S0047-259X(03)00096-4

Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Ann. Statist.* **40** 1024–1060. MR2985942 https://doi.org/10.1214/12-AOS989

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. MR3349498 https://doi.org/10.1016/j.jmva.2015.04.006

Ledoit, O. and Wolf, M. (2017). Numerical implementation of the QuEST function. *Comput. Statist. Data Anal.* **115** 199–223. MR3683138 https://doi.org/10.1016/j.csda.2017.06.004

Ledoit, O. and Wolf, M. (2018). Optimal estimation of a large-dimensional covariance matrix under Stein's loss. *Bernoulli* **24** 3791–3832. MR3788189 https://doi.org/10.3150/17-BEJ979

Ledoit, O. and Wolf, M. (2020). Analytical nonlinear shrinkage of large-dimensional covariance matrices. *Ann. Statist.* **48** 3043–3065. MR4152634 https://doi.org/10.1214/19-AOS1921

Ledoit, O. and Wolf, M.(2022). Supplement to "Quadratic shrinkage for large covariance matrices." https://doi.org/10.3150/20-BEJ1315SUPP

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

Loh, W.-L. (1991). Estimating covariance matrices. *Ann. Statist.* **19** 283–296. MR1091851 https://doi.org/10.1214/aos/1176347982

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

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

Markowitz, H. (1952). Portfolio Selection. *Journal of Finance* **7** 77–91.

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 https://doi.org/10.1007/3-540-31272-2_17

Nagao, H. (1973). On some test criteria for covariance matrix. *Ann. Statist.* **1** 700–709. MR0339405

Pal, N. (1993). Estimating the normal dispersion matrix and the precision matrix from a decision-theoretic point of view: A review. *Statist. Papers* **34** 1–26. MR1221520 https://doi.org/10.1007/BF02925524

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

Rajaratnam, B. and Vincenzi, D. (2016a). A theoretical study of Stein's covariance estimator. *Biometrika* **103** 653–666. MR3551790 https://doi.org/10.1093/biomet/asw030

Rajaratnam, B. and Vincenzi, D. (2016b). A note on covariance estimation in the unbiased estimator of risk framework. *J. Statist. Plann. Inference* **175** 25–39. MR3492762 https://doi.org/10.1016/j.jspi.2016.02.004

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

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* **4** Art. 32, 28. MR2183942 https://doi.org/10.2202/1544-6115.1175

Silverstein, J.W. (1995). Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. *J. Multivariate Anal.* **55** 331–339. MR1370408 https://doi.org/10.1006/jmva.1995.1083

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 https://doi.org/10.1006/jmva.1995.1051

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

Stein, C. (1986). Lectures on the theory of estimation of many parameters. *J. Math. Sci.* **34** 1373–1403.

Stoica, P., Li, J., Zhu, X. and Guerci, J.R. (2008). On using a priori knowledge in space-time adaptive processing. *IEEE Trans. Signal Process.* **56** 2598–2602. MR2516660 https://doi.org/10.1109/TSP.2007.914347

Tenenhaus, A. and Tenenhaus, M. (2011). Regularized generalized canonical correlation analysis. *Psychometrika* **76** 257–284. MR2788885 https://doi.org/10.1007/s11336-011-9206-8

Tsukuma, H. (2005). Estimating the inverse matrix of scale parameters in an elliptically contoured distribution. *J. Japan Statist. Soc.* **35** 21–39. MR2183498 https://doi.org/10.14490/jjss.35.21

Vakman, D. (1996). On the analytic signal, the Teager-Kaiser energy algorithm, and other methods for defining amplitude and frequency. *IEEE Transactions on Signal Processing* **44** 791–797.

Vidaurre, C., Krämer, N., Blankertz, B. and Schlögl, A. (2009). Time domain parameters as a feature for EEG-based brain–computer interfaces. *Neural Netw.* **22** 1313–1319.

Wigner, E.P. (1955). Characteristic vectors of bordered matrices with infinite dimensions. *Ann. of Math. (2)* **62** 548–564. MR0077805 https://doi.org/10.2307/1970079

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 https://doi.org/10.1111/j.1467-9868.2012.01049.x

Yang, R. and Berger, J.O. (1994). Estimation of a covariance matrix using the reference prior. *Ann. Statist.* **22** 1195–1211. MR1311972 https://doi.org/10.1214/aos/1176325625

Zhang, Y., Sun, D. and Zhang, D. (2009). Robust adaptive acoustic vector sensor beamforming using automated diagonal loading. *Appl. Acoust.* **70** 1029–1033.

Received November 2019 and revised December 2020