Journal of Multivariate Analysis 186 (2021) 104796

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Shrinkage estimation of large covariance matrices: Keep it simple, statistician?

Olivier Ledoit [a,b], Michael Wolf [a,*]

[a] Department of Economics, University of Zurich, Switzerland
[b] AlphaCrest Capital Management, NY, USA

ARTICLE INFO

Article history:
Received 12 October 2020
Received in revised form 14 July 2021
Accepted 14 July 2021
Available online 20 August 2021

AMS 2020 subject classifications:
primary 62H12
secondary 62G20
15A52

Keywords:
Large-dimensional asymptotics
Random matrix theory
Rotation equivariance

ABSTRACT

Under rotation-equivariant decision theory, sample covariance matrix eigenvalues can be optimally shrunk by recombining sample eigenvectors with a (potentially nonlinear) function of the unobservable population covariance matrix. The optimal shape of this function reflects the loss/risk that is to be minimized. We solve the problem of optimal covariance matrix estimation under a variety of loss functions motivated by statistical precedent, probability theory, and differential geometry. A key ingredient of our nonlinear shrinkage methodology is a new estimator of the angle between sample and population eigenvectors, without making strong assumptions on the population eigenvalues. We also introduce a broad family of covariance matrix estimators that can handle all regular functional transformations of the population covariance matrix under large-dimensional asymptotics. In addition, we compare via Monte Carlo simulations our methodology to two simpler ones from the literature, linear shrinkage and shrinkage based on the spiked covariance model.

1. Introduction

Ever since Charles Stein [30] proved that the usual estimator of the mean (vector) is inadmissible in dimensions greater than three, decision theory has taken the edge over likelihood maximization in multivariate statistics. This leaves open the question of which loss function to minimize in a practical application. In this respect, the more loss functions available the better, as different researchers may pursue different goals. Regarding the second moments, that is, covariance matrix estimation, six loss functions have been investigated so far within the framework of large-dimensional asymptotics by [8,22], yielding a grand total of three different optimal nonlinear shrinkage formulas.

This paper delivers the technology to double the number of loss functions that can be handled from 6 to 12, without making strict assumptions. The six new loss functions considered are potentially attractive to applied researchers, as they have been promoted before by statisticians for decision-theoretical estimation of the covariance matrix. In order to achieve this degree of generality, we identify a formula from random matrix theory (RMT) that enables us to develop a new estimator of the angle of any sample eigenvector with any population eigenvector, in the large-dimensional limit. Using this new technique opens the door to addressing a large set of loss functions that were previously unattainable within the framework of large-dimensional asymptotics with the techniques of [22]: In addition to the six specific new loss functions considered, we can also handle two infinite general families of loss functions based on all regular transformations of the population covariance matrix.

* Corresponding author.
E-mail address: michael.wolf@econ.uzh.ch (M. Wolf).

https://doi.org/10.1016/j.jmva.2021.104796

Before starting to develop our methodology, it will be useful to give a brief review of the relevant literature. Likelihood maximization has done wonders for statistics in general; however, in the particular context of multivariate statistics when the number of parameters to be estimated is large, it tends to overfit in-sample data, at the expense of good out-of-sample performance. In reaction to that, decision theory favors estimators that perform well out-of-sample with respect to some given loss function. These estimators critically depend on the loss function selected by the end-user.

For covariance matrix estimation, we place ourselves firmly within the paradigm pioneered by [32,33]: (i) no assumption on the eigenvalues of the population covariance matrix apart from positive definiteness; (ii) equivariance with respect to rotation of the original orthonormal basis of variables; and (iii) full flexibility to modify the eigenvalues of the sample covariance matrix as deemed necessary.

This is a tall order, and even Stein's finite-sample mathematical prowesses achieved limited progress. It was only after cross-pollination from RMT, a field originated by Nobel Prize-winning physicist Eugene Wigner [34], and specifically the notion of large-dimensional asymptotics, that conclusive strides forward could be made. Charles Stein himself was well aware, as early as 1969, of the potential of large-dimensional asymptotics to unlock the multivariate application problems that preoccupied him [31, pp. 79–81]. However, he left some work on the table for his intellectual successors in this respect.

There are currently three 'simplified' large-dimensional asymptotic strands of literature that fall short of Stein's ambitious program in one way or another. Sparsity [3] violates point (ii) because it assumes a priori knowledge of a specific orthonormal basis where (unlike for most other bases) the proportion of covariances equal to zero approaches 100%. Linear shrinkage [18] violates point (iii) because it can only modify the eigenvalues of the sample covariance matrix through a linear transformation. The spiked covariance model of [13] violates point (i) because it assumes that all population eigenvalues are equal to each other, except for a vanishingly small proportion of them (called 'spikes').

By contrast, the present paper inscribes itself in a strand of literature called nonlinear shrinkage [19,20,22] which does not compromise on any of these three points, and so remains in line with Stein's original ambitious paradigm. A key ingredient is consistent estimation of the eigenvalues of the population covariance matrix. This was not even deemed possible until [7] proved otherwise. Since then, it has been more of a discussion of which estimation scheme to use, such as [7]'s own numerical procedure or a more modern approach based on supersymmetry [14]; in this paper, we use the QuEST function of [20].

In recent related work, the spiked covariance model of [13] has been used by [5] to derive shrinkage covariance matrix estimators for a ménagerie of 26 different loss functions. [5] promote the spiked model because, as they state in their Section 10: "the simple shrinkage rules we propose here may be more likely to be applied correctly in practice, and to work as expected, even in relatively small sample sizes". It is, therefore, of interest to study whether our 'more complicated' nonlinear shrinkage rules actually lead to improved performance or whether applied researchers are just as well served by the rules of [5] according to their implicitly alluded to KISS (Keep it simple, statistician!) principle.

The remainder of this paper is organized as follows. Section 2 presents an intuitively understandable analysis in finite samples. Section 3 defines the large-dimensional asymptotics under which our results are derived. Section 4 investigates a wide variety of loss functions and, for each one, finds a bona fide covariance matrix estimator that is asymptotically optimal. Section 5 extends the analysis to the challenging yet empirically relevant case when the dimension exceeds the sample size. Section 6 presents Monte Carlo simulations. Section 7 concludes. An online supplementary material collects various mathematical results to keep the presentation of the paper compact, and also deals with the case of demeaning the data.

2. Analysis in finite samples

2.1. Basic setup

Assumption 1. Y is an $n \times p$ matrix of n independent and identically distributed (i.i.d.) observations on a system of $p < n$ random variables with mean zero and positive definite covariance matrix Σ with eigenvalues $(\tau_1, \ldots, \tau_p)$, sorted in nondecreasing order without loss of generality (w.l.o.g.), and corresponding eigenvectors $(v_1, \ldots, v_p)$.

(The case $p > n$ is treated in Section 5.)

The sample covariance matrix is $S := Y^\top Y/n$. Its spectral decomposition is $S =: U \Lambda U^\top$, where Λ is a diagonal matrix and U is orthogonal. Let $\Lambda =: \mathrm{Diag}(\boldsymbol{\lambda})$ where $\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_p)^\top$, with the eigenvalues again sorted in nondecreasing order w.l.o.g. The ith sample eigenvector is u_i, the ith column vector of U, so that $S = \sum_{i=1}^p \lambda_i \cdot u_i u_i^\top$. Note that it holds similarly $\Sigma = \sum_{i=1}^p \tau_i \cdot v_i v_i^\top$.

Definition 1. We consider rotation-equivariant covariance matrix estimators of the type $\widetilde{S} := U \widetilde{D} U^\top$, where $\widetilde{D}$ is a diagonal matrix: $\widetilde{D} := \mathrm{Diag}(\widetilde{d}_1, \ldots, \widetilde{d}_p)$.

This class assumes no a priori information about the orientation of the orthonormal basis of the (unobservable) population covariance-matrix eigenvectors; this is different from the sparsity literature, which requires a priori knowledge

Table 1
Existing set of finite-sample optimal (FSOPT) nonlinear shrinkage formulas. The first column gives the name of the loss function; the second column gives the corresponding stylized formula; the third column gives a reference for the loss function; and the fourth column gives the shrinkage formula for $\widetilde{d}_i$, that is, the replacement for the ith sample eigenvalue.

Loss function	Stylized Formula	Reference	Shrinkage
Frobenius	$\left\|\widetilde{S}-\Sigma\right\|_{\mathrm{F}}$	[24]	$u_i^\top \Sigma u_i$
Inverse Stein	$\mathrm{Tr}\left[\widetilde{S}^{-1}\Sigma\right]-\log\left[\det\left(\widetilde{S}^{-1}\Sigma\right)\right]$	[9]	$u_i^\top \Sigma u_i$
Minimum Variance	$\mathrm{Tr}\left[\widetilde{S}^{-1}\Sigma\widetilde{S}^{-1}\right]\Big/\left(\mathrm{Tr}\left[\widetilde{S}^{-1}\right]\right)^2$	[8]	$u_i^\top \Sigma u_i$
Stein	$\mathrm{Tr}\left[\widetilde{S}\Sigma^{-1}\right]-\log\left[\det\left(\widetilde{S}\Sigma^{-1}\right)\right]$	[12]	$\frac{1}{u_i^\top \Sigma^{-1}u_i}$
Inverse Frobenius	$\left\|\widetilde{S}^{-1}-\Sigma^{-1}\right\|_{\mathrm{F}}$	[10]	$\frac{1}{u_i^\top \Sigma^{-1}u_i}$
Symmetrized Stein	$\mathrm{Tr}\left[\widetilde{S}\Sigma^{-1}+\widetilde{S}^{-1}\Sigma\right]$	[15]	$\sqrt{\frac{u_i^\top \Sigma u_i}{u_i^\top \Sigma^{-1}u_i}}$

Table 2
Two more loss functions leading to existing nonlinear shrinkage formulas. The first column gives the name of the loss function; the second column gives the corresponding stylized formula; the third column gives a reference for the loss function; and the fourth column gives the shrinkage formula for $\widetilde{d}_i$, that is, the replacement for the ith sample eigenvalue.

Loss function	Stylized formula	Reference	Shrinkage
Weighted Frobenius	$\mathrm{Tr}\left[\left(\widetilde{S}-\Sigma\right)^2\Sigma^{-1}\right]$	Sharma and Krishnamoorthy ([27])	$\frac{1}{u_i^\top \Sigma^{-1}u_i}$
Disutility	$\mathrm{Tr}\left[\left(\widetilde{S}^{-1}-\Sigma^{-1}\right)^2\Sigma\right]$	Supplementary material A	$u_i^\top \Sigma u_i$

of an orthonormal basis in which most covariances are (close to) zero. For many loss functions, there exists a finite-sample optimal (FSOPT) estimator in this class of the form

$$\widetilde{S} \coloneqq \sum_{i=1}^{p} \widetilde{d}_i \cdot u_i u_i^\top \ , \qquad \widetilde{d}_i \coloneqq \gamma^{-1}\left[u_i^\top \gamma(\Sigma) u_i\right] \ , \qquad i \in \{1,\ldots,p\} \ , \tag{1}$$

where γ denotes some smooth invertible function mapping of $(0,+\infty)$ onto $\mathbb{R}$. Here, as is standard, applying a univariate function γ to a diagonalizable positive-definite matrix means preserving its eigenvectors and applying γ to each eigenvalue individually; for example, $\log(\Sigma) \coloneqq \sum_{i=1}^{p} \log(\tau_i)\cdot v_i v_i^\top$. Furthermore, γ^{-1} denotes the inverse function, so for example if $\gamma(x)=x^3$ then $\gamma^{-1}(x)$ is equal to $x^{1/3}$, not x^{-3}

Remark 1. $\widetilde{S}$ in (1) is not feasible in practice, so the term "estimator" can be considered a slight abuse of terminology in this context.

Remark 2. To simplify the notation, and in line with the related literature, we assume throughout the paper that all variables have mean zero. Supplementary material C deals with the case when this assumption is not (known to be) true.

2.2. A brief summary of known results on nonlinear shrinkage

So far, only six loss functions have been solved in the very general rotation-equivariant framework of Assumption 1 and Definition 1. In the second column of Table 1, the loss functions are streamlined for readability; the actual ones could be squared and have various constants added or multiplied in ways that are irrelevant to estimator optimality. The way to read the fourth column is that the ith sample eigenvalue $\lambda_i = u_i^\top S u_i$, $i \in \{1,\ldots,p\}$, should be replaced by the quantity in the fourth column, optimally with respect to the same-row loss function, in finite samples: so it is the optimally 'shrunk' eigenvalue. We use the standard notation for the Frobenius norm of M, a square matrix: $\|M\|_{\mathrm{F}} \coloneqq \sqrt{\mathrm{Tr}[MM^\top]}$.

Table 1 shows that the six loss functions really only yield three different nonlinear shrinkage formulas. The first two are of the type (1) with $\gamma(x)=x$ and $\gamma(x)=1/x$, respectively, and the third one is simply their geometric mean.

2.3. Additional loss functions

The easiest way to start this investigation is to look for different loss functions that give rise to the same nonlinear shrinkage formulas as the ones in Section 2.2. Table 2 presents two of them.

The second loss function, called Disutility, is new. It is derived from the loss of [27] in the same way that the Inverse Frobenius loss is derived from the Frobenius loss, or that the Inverse Stein's loss of [9] is derived from the original Stein's loss: by substituting the covariance matrix with its inverse, the precision matrix. At the same time, it has a more interesting justification as minus the quadratic utility function of [25] in large dimensions, as argued in Supplementary material A (hence the name disutility). It is a close cousin of the Minimum Variance loss function, with a tighter grip on the scale of the estimator. Reassuringly, both of them give rise to the same optimal nonlinear shrinkage formula.

There are three interlocking reasons for bringing up these loss functions, even though they fall back on the known estimators of Section 2.2. First, to avoid the well-known 'file-drawer problem' (also called publication bias), whereby results that are deemed less interesting remain unpublished. Second, some applied researcher may well look at one of these two loss functions and recognize that it suits his or her objective perfectly, in which case it does not matter whether the shrinkage formula is old or new. Third, in the end, the choice of estimator is a choice of shrinkage formula, and the best way to know what a specific shrinkage really means is to list as many loss functions as possible that lead to it.

2.4. New shrinkage formulas

The main point of the paper is to go beyond the two cases $\gamma(x) = x^{\pm 1}$ and thereby to study other functions of the population covariance matrix (through the prism of sample eigenvectors). We introduce four more: $\sqrt{x}$, $\log(x)$, x^2, and $1/x^2$. Hence, we triple the number of functions that can be utilized for this purpose, from two to six. We could have introduced as many new functions as we wanted, but this should be enough to make the point. Nor is this frivolous or arbitrary: these four functional transformations arise naturally in the study of four well-regarded loss functions that have remained as open problems. In what follows, the symbol $\mathbb{I}$ denotes a conformable identity matrix.

Log-Euclidian It is defined as the Euclidian distance on the logarithm of the manifold of symmetric positive-definite matrices, hence the name. It is a close cousin of the geodesic distance on the smooth Riemannian manifold of positive-definite matrices. It has essentially the same properties, but is much more tractable for statistical applications. In particular, it is invariant with respect to matrix inversion, so eigenvalues close to zero are treated like eigenvalues close to infinity.

Fréchet The Fréchet discrepancy, named after the French mathematician Maurice Fréchet (1878–1973), is originally a measure of distance between two probability distributions. In the multivariate normal case, it directly implies a notion of distance between any two symmetric positive-definite matrices. Intuitively, we should think of it as a measure of 'how far apart' are the distributions that these two covariance matrices generate.

Quadratic This is a recent variant of the quadratic-type loss function that can be traced back to pioneers in the field such as Section 2.2.4 of [26] and loss function L_2 of [11]. Its signature is that it promotes accuracy in the direction of the smallest principal components of the population covariance matrix.

Inverse Quadratic Same as above, but with the inverse sample covariance matrix. Mechanically, it promotes accuracy in the direction of the largest principal components of the population covariance matrix.

The logarithm and the square root are directly embedded into the first two shrinkage formulas (Log-Euclidian and Fréchet), but the square and inverse-square functions only appear in the last two loss formulas as part of combinations, echoing what happened with the Symmetrized Stein's loss. Proof that the loss functions in the second column of the tables give rise to the FSOPT formulas in the fourth column can be found in Supplementary material B.

These seven nonlinear shrinkage formulas give rather different results. Researchers may wonder how they compare to each other. One interesting mathematical observation is that they do not cross, but one is always above (or below) the other across the whole spectrum — with the sole exception of Symmetrized Stein vs. Log-Euclidian shrinkage. The following proposition reveals the ordering.

Proposition 1. *Under Assumption 1, with probability one, $i \in \{1, \ldots, p\}$.*

$$\frac{u_i^\top \Sigma^2 u_i}{u_i^\top \Sigma u_i} \geq u_i^\top \Sigma u_i \geq \left(u_i^\top \sqrt{\Sigma} u_i\right)^2 \geq \exp\left[u_i^\top \log(\Sigma)\, u_i\right] \geq \frac{1}{u_i^\top \Sigma^{-1} u_i} \geq \frac{u_i^\top \Sigma^{-1} u_i}{u_i^\top \Sigma^{-2} u_i},$$

$$\left(u_i^\top \sqrt{\Sigma} u_i\right)^2 \geq \sqrt{\frac{u_i^\top \Sigma u_i}{u_i^\top \Sigma^{-1} u_i}} \geq \frac{1}{u_i^\top \Sigma^{-1} u_i}.$$

Proof. Follows from Jensen's inequality and the Cauchy–Schwarz inequality once we remark that $u_i^\top \gamma(\Sigma) u_i = \sum_{j=1}^p \gamma(\tau_i) \cdot \left(u_i^\top v_j\right)^2$ for $\gamma(x) = x$, $1/x$, x^2, $1/x^2$, $\sqrt{x}$, or $\log(x)$, and that $\sum_{j=1}^p \left(u_i^\top v_j\right)^2 = 1$, $i \in \{1, \ldots, p\}$.

Table 3
New set of finite-sample optimal (FSOPT) nonlinear shrinkage formulas. The first column gives the name of the loss function; the second column gives the corresponding formula; the third column gives a reference for the loss function; and the fourth column gives the shrinkage formula for $\widetilde{d}_i$, that is, the replacement for the ith sample eigenvalue.

Loss function	Formula	Reference	Shrinkage
Log-Euclidian	$\left\|\log(\widetilde{S}) - \log(\Sigma)\right\|_{\mathrm{F}}$	[1]	$\exp[u_i^\top \log(\Sigma) u_i]$
Fréchet	$\left\|\widetilde{S}^{1/2} - \Sigma^{1/2}\right\|_{\mathrm{F}}$	[6]	$\left(u_i^\top \Sigma^{1/2} u_i\right)^2$
Quadratic	$\left\|\Sigma^{-1}\widetilde{S} - \mathbb{I}\right\|_{\mathrm{F}}$	$L^{F,3}$ in [5]	$\frac{u_i^\top \Sigma^{-1} u_i}{u_i^\top \Sigma^{-2} u_i}$
Inverse Quadratic	$\left\|\widetilde{S}^{-1}\Sigma - \mathbb{I}\right\|_{\mathrm{F}}$	$L^{F,4}$ in [5]	$\frac{u_i^\top \Sigma^{2} u_i}{u_i^\top \Sigma u_i}$

2.5. Preview of general result

FSOPT estimators of the form (1) cannot be used directly because they depend on the population covariance matrix Σ, which is unobservable. So it stands to reason to ask: How is it even possible that this approach leads anywhere? First of all, note that we do not need to estimate all $p(p+1)/2$ entries of the symmetric matrix Σ, we only need p quantities: $u_i^\top \gamma(\Sigma) u_i$, $i \in \{1, \ldots, p\}$, which is much more manageable. When the matrix dimension p is large, it is possible to approximate these quantities by the general formula

$$u_i^\top \gamma(\Sigma) u_i \approx \frac{1}{p} \sum_{j=1}^{p} \gamma(\widehat{\tau}_j) \cdot \left\{ \frac{\frac{p}{n}\lambda_i \widehat{\tau}_j}{\left|\widehat{\tau}_j \left[1 - \frac{p}{n} - \frac{p}{n}\lambda_i \breve{m}_{n,p}^{\widehat{\tau}}(\lambda_i)\right] - \lambda_i\right|^2} \right\}, \tag{2}$$

where $\widehat{\boldsymbol{\tau}} := (\widehat{\tau}_1, \ldots, \widehat{\tau}_p)^\top$ is an estimator of the population eigenvalues, and $\breve{m}_{n,p}^{\widehat{\tau}}(\cdot)$ is the complex-valued function of real argument due to Section 2 of [20]. Formula (2) generates bona fide covariance matrix estimators of the type (1) for all the loss functions in Table 3 by setting $\gamma(x)$ equal to $\log(x)$, $\sqrt{x}$, x^{-2}, or x^2. Given that $u_i^\top \gamma(\Sigma) u_i = \frac{1}{p}\sum_{j=1}^{p} \gamma(\tau_j) \cdot \left\{p(u_i^\top v_j)^2\right\}$, the term between curly brackets in (2) is simply an estimator of the dimension-normalized squared dot product of the ith sample eigenvector with the jth population eigenvector.

3. Analysis under large-dimensional asymptotics

We now move on to formally establishing that plugging the approximation (2) into the generic nonlinear shrinkage formula (1) yields optimal rotation-equivariant covariance matrix estimators under large-dimensional asymptotics with respect to the loss functions listed. First of all, to make the paper self-contained, we need to restate some sets of assumptions that have been used a number of times before. We shall do so in a condensed fashion; any unfamiliar reader interested in getting more background information should refer to earlier work such as, for example, Section 3.1 of [22] and the references therein.

In a nutshell: The dimension p goes to infinity along with the sample size n, their ratio p/n converges to some limit $c \in (0, 1)$, and we seek to asymptotically optimize the way to nonlinearly shrink sample eigenvalues. Also, from now on, all dimension-dependent objects are subscripted by the sample size n.

3.1. Large-dimensional asymptotic framework

Assumption 2 (*Dimension*)**.** Let n denote the sample size and $p := p(n)$ the number of variables. It is assumed that the ratio p/n converges, as $n \to \infty$, to a limit $c \in (0, 1)$ called the limiting concentration (ratio). Furthermore, there exists a compact interval included in $(0, 1)$ that contains p/n for all n large enough.

Assumption 3 (*Population Covariance Matrix*)**.**

a. The $p \times p$ population covariance matrix Σ_n is nonrandom symmetric positive-definite.
b. Let $\boldsymbol{\tau}_n := (\tau_{n,1}, \ldots, \tau_{n,p})^\top$ denote a system of eigenvalues of Σ_n, and H_n their empirical distribution function (e.d.f.): $H_n(x) := \sum_{i=1}^{p} \mathbb{1}_{[\tau_{n,i}, +\infty)}(x)/p$, where $\mathbb{1}$ denotes the indicator function of a set. It is assumed that H_n converges weakly to some limit law H, called the limiting spectral distribution (function).
c. Supp(H), the support of H, is the union of a finite number of closed intervals in $(0, +\infty)$.
d. There exists a compact interval $[\underline{h}, \overline{h}] \subset (0, \infty)$ that contains $\left\{\tau_{n,1}, \ldots, \tau_{n,p}\right\}$ for all n large enough.

This assumption includes the spiked covariance model of [13] as a special case in which the limiting population spectral distribution H is assumed to be a point mass.

Assumption 4 (*Data Generating Process*)**.** X_n is an $n \times p$ matrix of i.i.d. random variables with mean zero, variance one, and finite 12th moment. The matrix of observations is $Y_n := X_n\sqrt{\Sigma_n}$. Neither $\sqrt{\Sigma_n}$ nor X_n are observed on their own: only Y_n is observed.

This assumption includes the spiked covariance model of [13] as a special case in which the variates are assumed to be normal.

Remark 3 (*Moment Condition*)**.** The existence of a finite 12th moment is assumed to prove certain mathematical results using the methodology of [17]. However, Monte Carlo studies in [19,20] indicate that this assumption is not needed in practice and can be replaced with the existence of a finite fourth moment. This is a generic requirement which does not depend on any particular loss function.

The sample covariance matrix is defined as $S_n := n^{-1}Y_n^\top Y_n = n^{-1}\sqrt{\Sigma_n}X_n^\top X_n\sqrt{\Sigma_n}$. It admits a spectral decomposition $S_n =: U_n\Lambda_n U_n^\top$, where Λ_n is a diagonal matrix, and U_n is an orthogonal matrix: $U_nU_n^\top = U_n^\top U_n = \mathbb{I}_n$, where $\mathbb{I}_n$ (in slight abuse of notation) denotes the identity matrix of dimension $p \times p$. Let $\Lambda_n := \mathrm{Diag}(\boldsymbol{\lambda}_n)$ where $\boldsymbol{\lambda}_n := (\lambda_{n,1}, \ldots, \lambda_{n,p})^\top$. We assume w.l.o.g. that the sample eigenvalues are sorted in increasing order: $\lambda_{n,1} \leq \lambda_{n,2} \leq \cdots \leq \lambda_{n,p}$. Correspondingly, the ith sample eigenvector is $u_{n,i}$, the ith column vector of U_n. Under Assumptions 2–4, the e.d.f. of the sample eigenvalues $F_n(x) := \sum_{i=1}^p \mathbb{1}_{[\lambda_{n,i},+\infty)}(x)/p$ converges almost surely to a nondeterministic cumulative distribution function F that depends only on H and c:

$$F_n(x) \xrightarrow{\text{a.s.}} F(x) \qquad \forall x \in (0, +\infty)\,.$$

How to go from (H, c) to F is determined by the following equation, which is due to [28]: For all z in $\mathbb{C}^+$, the half-plane of complex numbers with strictly positive imaginary part, $m := m_F(z)$ is the unique solution in the set $\left\{m \in \mathbb{C} : -\frac{1-c}{z} + cm \in \mathbb{C}^+\right\}$ to the equation

$$m = \int \frac{1}{\tau\big[1 - c - c\,z\,m\big] - z}\,dH(\tau)\,, \tag{3}$$

where m_F denotes the Stieltjes transform of F, whose standard definition is:

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

The Stieltjes transform admits a well-known inversion formula:

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

if G is continuous at both a and b, where Im denotes the imaginary part of a complex number. Although the Stieltjes transform of F, m_F, is a function whose domain is the upper half of the complex plane, it admits an extension to the real line, since [29] show that: $\forall x \in (0, +\infty)$, $\lim_{z \in \mathbb{C}^+ \to x} m_F(z) =: \breve{m}_F(x)$ exists and is continuous. The imaginary part of $\breve{m}_F$ is the derivative of F, up to rescaling by π; therefore, (3) enables us to pin down the location of the sample eigenvalues, a fact exploited by the QuEST function; see Section 3.2. Furthermore, the support of the limiting distribution of the sample eigenvalue, $\mathrm{Supp}(F)$, is the union of a finite number $\kappa \geq 1$ of compact intervals: $\mathrm{Supp}(F) = \bigcup_{k=1}^{\kappa}[a_k, b_k]$, where $0 < a_1 < b_1 < \cdots < a_\kappa < b_\kappa < \infty$.

Definition 2 (*Rotation-Equivariant Estimators*)**.** We consider covariance matrix estimators of the type $\widetilde{S}_n := U_n\widetilde{D}_nU_n^\top$, where $\widetilde{D}_n$ is a diagonal matrix: $\widetilde{D}_n := \mathrm{Diag}(\widetilde{\varphi}_n(\lambda_{n,1})\ldots, \widetilde{\varphi}_n(\lambda_{n,p}))$, and $\widetilde{\varphi}_n$ is a (possibly random) real univariate function which can depend on S_n.

Assumption 5 (*Nonlinear Shrinkage Function*)**.** We assume that there exists a nonrandom real univariate function $\widetilde{\varphi}$ defined on $\mathrm{Supp}(F)$ and continuously differentiable on $\bigcup_{k=1}^{\kappa}[a_k, b_k]$ such that $\widetilde{\varphi}_n(x) \xrightarrow{\text{a.s.}} \widetilde{\varphi}(x)$ for all $x \in \mathrm{Supp}(F)$. Furthermore, this convergence is uniform over $x \in \bigcup_{k=1}^{\kappa}[a_k + \eta, b_k - \eta]$, for any small $\eta > 0$. Finally, for any small $\eta > 0$, there exists a finite nonrandom constant $\widetilde{K}$ such that almost surely, over the set $x \in \bigcup_{k=1}^{\kappa}[a_k - \eta, b_k + \eta]$, $|\widetilde{\varphi}_n(x)|$ is uniformly bounded by $\widetilde{K}$, for all n large enough.

3.2. The QuEST function

Once again, to make the paper self-contained, we need to restate the definition of a key mathematical object called the QuEST (quantized eigenvalues sampling transform) function. We shall do so in condensed fashion; the interested reader is referred to [20,21] for full background information.

In a nutshell: QuEST is a multivariate deterministic function mapping population eigenvalues into sample eigenvalues, valid asymptotically as p and n go to infinity together.

Definition 3 (*QuEST*). For any given n and p, $Q_{n,p}$ maps $\mathbf{t} := \left(t_1, \ldots, t_p\right)^\top \in [0, +\infty)^p$ into

$$Q_{n,p}(\mathbf{t}) := \left(q^1_{n,p}(\mathbf{t}), \ldots, q^p_{n,p}(\mathbf{t})\right)^\top, \quad \text{where} \quad q^i_{n,p}(\mathbf{t}) := p \int_{(i-1)/p}^{i/p} \left(F^{\mathbf{t}}_{n,p}\right)^{-1}(u)\, du \,,$$

$$\left(F^{\mathbf{t}}_{n,p}\right)^{-1} \text{is the inverse function of} \quad F^{\mathbf{t}}_{n,p}(v) := \frac{1}{\pi} \int_{-\infty}^{v} \operatorname{Im}\left[\breve{m}^{\mathbf{t}}_{n,p}(x)\right] dx \,,$$

and, for all x in $\mathbb{R}$, $\breve{m}^{\mathbf{t}}_{n,p}(x)$ is the unique solution $m \in \mathbb{C}^+$ to the fundamental equation:

$$m = \frac{1}{p} \sum_{j=1}^{p} \frac{1}{t_j \left(1 - \frac{p}{n} - \frac{p}{n} x\, m\right) - x} \,. \tag{4}$$

Theorem 1 (*[20]*). *Suppose Assumptions 2–4 are satisfied. Define*

$$\widehat{\boldsymbol{\tau}}_n := \operatorname*{argmin}_{\mathbf{t} \in (0,+\infty)^p} \frac{1}{p} \sum_{i=1}^{p} \left[q^i_{n,p}(\mathbf{t}) - \lambda_{n,i}\right]^2 , \tag{5}$$

where $Q_{n,p}(\mathbf{t})$ is the QuEST function from Definition 3; both $\widehat{\boldsymbol{\tau}}_n$ and $\boldsymbol{\lambda}_n$ are assumed sorted in nondecreasing order. Let $\widehat{\tau}_{n,j}$ denote the jth entry of $\widehat{\boldsymbol{\tau}}_n$ ($j = 1, \ldots, p$), and let $\boldsymbol{\tau}_n := (\tau_{n,1}, \ldots, \tau_{n,p})^\top$ denote the population covariance matrix eigenvalues sorted in nondecreasing order w.l.o.g. Then

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

The function $\breve{m}^{\widehat{\boldsymbol{\tau}}_n}_{n,p}$ featured in the approximation (2) is a by-product of the QuEST function constructed by combining (4)–(5). It estimates the complex-valued deterministic function of real argument $\breve{m}_F$.

3.3. Dot product of population eigenvalues with sample eigenvalues

Of much importance in this paper is the random bivariate cumulative distribution function

$$\forall x, t \in \mathbb{R} \qquad \Theta_n(x, t) := \frac{1}{p} \sum_{i=1}^{p} \sum_{j=1}^{p} (u^\top_{n,i} v_{n,j})^2\, \mathbb{1}_{[\lambda_{n,i}, +\infty)}(x) \cdot \mathbb{1}_{[\tau_{n,j}, +\infty)}(t) \,, \tag{6}$$

which was first introduced in (6) of [17] under the notation Φ_N. From Θ_n we can extract precise information about the relationship between sample and population eigenvectors. In theory, the dot product $u^\top_{n,i} v_{n,j}$would be something worth looking at. However, the sign is irrelevant, so we focus on the square $(u^\top_{n,i} v_{n,j})^2$ instead. Even then, we have to bear in mind that we operate under large-dimensional asymptotics, so all quantities need to be normalized by the ever-increasing matrix dimension p in appropriate fashion. In this particular instance, $(u^\top_{n,i} v_{n,j})^2$ vanishes at the speed $1/p$, as can be seen from the following identities:

$$\frac{1}{p^2} \sum_{i=1}^{p} \sum_{j=1}^{p} (u^\top_{n,i} v_{n,j})^2 = \frac{1}{p^2} \sum_{i=1}^{p} u^\top_{n,i} \left(\sum_{j=1}^{p} v_{n,j} v^\top_{n,j}\right) u_{n,i} = \frac{1}{p^2} \sum_{i=1}^{p} u^\top_{n,i} u_{n,i} = \frac{1}{p} \,.$$

Therefore, it is more convenient to study $p(u^\top_{n,i} v_{n,j})^2$ instead. The average of the quantities of interest $p(u^\top_{n,i} v_{n,j})^2$ over the sample (respectively population) eigenvectors associated with the sample (respectively population) eigenvalues lying in the interval $[\underline{\lambda}, \overline{\lambda}]$ (respectively $[\underline{\tau}, \overline{\tau}]$) is equal to

$$\frac{\sum_{i=1}^{p} \sum_{j=1}^{p} p(u^\top_{n,i} v_{n,j})^2 \mathbb{1}_{[\underline{\lambda}, \overline{\lambda}]}(\lambda_{n,i}) \cdot \mathbb{1}_{[\underline{\tau}, \overline{\tau}]}(\tau_{n,j})}{\sum_{i=1}^{p} \sum_{j=1}^{p} \mathbb{1}_{[\underline{\tau}, \overline{\tau}]}(\tau_{n,j})} = \frac{\Theta_n(\overline{\lambda}, \overline{\tau}) - \Theta_n(\underline{\lambda}, \overline{\tau}) - \Theta_n(\overline{\lambda}, \underline{\tau}) + \Theta_n(\underline{\lambda}, \underline{\tau})}{\left[F_n(\overline{\lambda}) - F_n(\underline{\lambda})\right] \cdot \left[H_n(\overline{\tau}) - H_n(\underline{\tau})\right]} \,.$$

Thus, the object of interest is the Radon–Nikodym derivative of (the limit of) $\Theta_n(x, t)$ with respect to the cross-product $F(x)\, H(t)$; which is exactly what (7) delivers.

Theorem 2 (*[17]*). *Under Assumptions 2–4, $\forall \lambda, \tau \in \mathbb{R}$, $\Theta_n(\lambda, \tau)$ converges almost surely to some nonrandom bivariate cumulative distribution function $\Theta(\lambda, \tau) := \int_{-\infty}^{\lambda} \int_{-\infty}^{\tau} \theta(x, t)\, dH(t)\, dF(x)$, where*

$$\forall x, t \in \mathbb{R} \qquad \theta(x, t) := \frac{cxt}{\left|t\left[1 - c - c\, x\, \breve{m}_F(x)\right] - x\right|^2} \,. \tag{7}$$

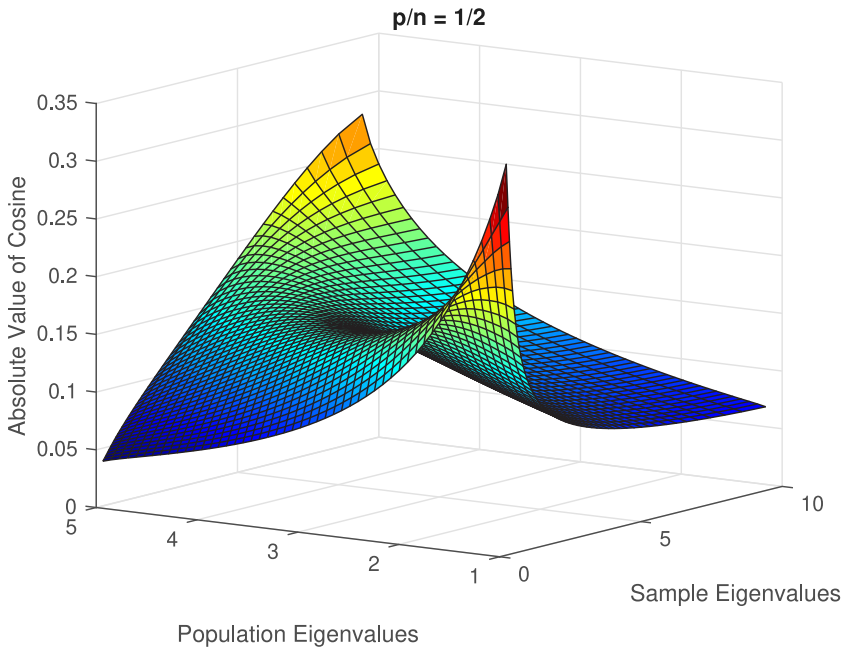

Fig. 1. Absolute value of the cosine of the angle between population and sample eigenvectors. On the horizontal axes, eigenvectors are indexed by their respective eigenvalues.

The Radon–Nikodym derivative $\theta(\lambda_{n,i}, \tau_{n,j})$ is 'essentially like' the squared dot product $p(u_{n,i}^\top v_{n,j})^2$ for large p and n. In order to operationalize (7), we need bona fide estimators for its ingredients, and they are provided by Section 3.2's QuEST function:

$$\widehat{\theta}_n(\lambda_{n,i}, \widehat{\tau}_{n,j}) := \frac{\frac{p}{n}\lambda_{n,i}\widehat{\tau}_{n,j}}{\left|\widehat{\tau}_{n,j}\left[1 - \frac{p}{n} - \frac{p}{n}\lambda_{n,i}\breve{m}_{n,p}^{\widehat{\tau}_n}(\lambda_{n,i})\right] - \lambda_{n,i}\right|^2}. \tag{8}$$

Although the expression may seem a bit unusual, it is just what comes out of RMT, and we should count ourselves lucky to have any closed-form solution at all. This 'luck' is first and foremost due to the pioneering efforts of probabilists who came before. If (3), (4), (7), and (8) appear to be descendants from each other, it is because they are. A graphical illustration in the case where the population eigenvalues are evenly spread in the interval $[1, 5]$, with concentration ratio $p/n = 0.5$, is given by Fig. 1.

One can see that the spread of sample eigenvalues is much wider: from 0.2 to 10.2. Top-ranked sample eigenvectors are more aligned with top-ranked population eigenvectors, and bottom-ranked sample eigenvectors are more aligned with bottom-ranked population eigenvectors. The overall pattern is complicated and can only be captured by the function θ of Theorem 2.

4. Asymptotically optimal nonlinear shrinkage estimators

The two loss functions from Table 2 are easy to handle using the techniques of [22]: The nonlinear shrinkage estimator they call $\widehat{S}_n^*$ is optimal with respect to the Weighted Frobenius loss under large-dimensional asymptotics; and the estimator they call $\widehat{S}_n^\circ$ is optimal with respect to the Disutility loss. These results are stated without proof, as they are just minor extensions of the arguments put forward by [22].

Regarding the loss functions of Table 3, they are vastly more challenging, and cannot be handled with existing techniques. Instead, they can only be handled by using the new technique of angle estimation introduced in Section 3.3 above, as we shall now proceed to demonstrate.

4.1. Four specific loss functions

We start with asymptotically optimal bona fide estimators based on Table 3.

Theorem 3 (*Log-Euclidian*). *For any estimator $\widetilde{S}_n$ in Definition 2, the Log-Euclidian loss*

$$\mathcal{L}_n^{LE}(\widetilde{S}_n, \Sigma_n) := \frac{1}{p}Tr\left[\left\{\log\left(\widetilde{S}_n\right) - \log\left(\Sigma_n\right)\right\}^2\right],$$

converges under Assumptions 2–5 almost surely to a deterministic limit that depends only on H, c, and $\widetilde{\varphi}$. This limit is minimized if $\widetilde{\varphi}_n(\lambda_{n,i})$ is equal to

$$\widehat{\varphi}_n^{LE}(\lambda_{n,i}) := \exp\left(\frac{1}{p}\sum_{j=1}^{p}\log(\widehat{\tau}_{n,j})\cdot\widehat{\theta}_n(\lambda_{n,i},\widehat{\tau}_{n,j})\right) , \tag{9}$$

where $\widehat{\boldsymbol{\tau}}_n = \left(\widehat{\tau}_{n,j}\right)_{j=1,\ldots,p}$ denotes the estimator of population covariance matrix eigenvalues in Theorem 1, and $\widehat{\theta}_n(\lambda_{n,i},\widehat{\tau}_{n,j})$ is the estimator of the (dimension-normalized) squared dot product of the ith sample eigenvector with the jth population eigenvector in (8). The resulting covariance matrix estimator is $\widehat{S}_n^{LE} := \sum_{i=1}^{p}\widehat{\varphi}_n^{LE}\left(\lambda_{n,i}\right)\cdot u_{n,i}u_{n,i}^{\top}$.

Proof. This theorem is a special case of Theorem 7 with $\gamma(x) = \log(x)$.

Theorem 4 (*Fréchet*). *The Fréchet loss $\mathcal{L}_n^{FR\acute{E}}(\widetilde{S}_n, \Sigma_n) := \left\|\widetilde{S}_n^{1/2} - \Sigma_n^{1/2}\right\|_F^2/p$ converges almost surely to a deterministic limit that is minimized if $\widetilde{\varphi}_n(\lambda_{n,i})$ is equal to*

$$\widehat{\varphi}_n^{FR\acute{E}}(\lambda_{n,i}) := \left(\frac{1}{p}\sum_{j=1}^{p}\sqrt{\widehat{\tau}_{n,j}}\cdot\widehat{\theta}_n(\lambda_{n,i},\widehat{\tau}_{n,j})\right)^2 . \tag{10}$$

The resulting covariance matrix estimator is $\widehat{S}_n^{FR\acute{E}} := \sum_{i=1}^{p}\widehat{\varphi}_n^{FR\acute{E}}\left(\lambda_{n,i}\right)\cdot u_{n,i}u_{n,i}^{\top}$.

Proof. This theorem is a special case of Theorem 7 with $\gamma(x) = \sqrt{x}$.

Theorem 5 (*Quadratic*). *The Quadratic loss $\mathcal{L}^{Q}\left(\widetilde{S}_n, \Sigma_n\right) := \left\|\Sigma_n^{-1}\widetilde{S}_n - \mathbb{I}_n\right\|_F^2/p$ converges almost surely to a deterministic limit that is minimized if $\widetilde{\varphi}_n(\lambda_{n,i})$ is equal to*

$$\widehat{\varphi}_n^{Q}(\lambda_{n,i}) := \frac{\frac{1}{p}\sum_{j=1}^{p}\frac{1}{\widehat{\tau}_{n,j}}\cdot\widehat{\theta}_n(\lambda_{n,i},\widehat{\tau}_{n,j})}{\frac{1}{p}\sum_{j=1}^{p}\frac{1}{\widehat{\tau}_{n,j}^2}\cdot\widehat{\theta}_n(\lambda_{n,i},\widehat{\tau}_{n,j})} . \tag{11}$$

The resulting covariance matrix estimator is $\widehat{S}_n^{Q} := \sum_{i=1}^{p}\widehat{\varphi}_n^{Q}\left(\lambda_{n,i}\right)\cdot u_{n,i}u_{n,i}^{\top}$.

Proof. To prove the theorem, we need to first state and prove some auxiliary results.

Proposition 2. *Under Assumptions 2–5,*

$$\mathcal{L}_n^{Q}(\widetilde{S}_n, \Sigma_n) \xrightarrow{a.s.} \sum_{k=1}^{\kappa}\int_{a_k}^{b_k}\int_{-\infty}^{+\infty}\left[\frac{\widetilde{\varphi}(x)^2}{t^2} - 2\frac{\widetilde{\varphi}(x)}{t} + 1\right]\cdot\theta(x,t)\,dH(t)\,dF(x) . \tag{12}$$

Proof. For simplicity, let us assume that the support of F is a single compact interval $[a, b] \subset (0, +\infty)$; the generalization to the case $\kappa > 1$ is trivial. From Supplementary material B.11 we have:

$$\begin{aligned}
\mathcal{L}_n^{Q}(\widetilde{S}_n, \Sigma_n) &= \frac{1}{p}\sum_{i=1}^{p}\widetilde{d}_{n,i}^2\cdot u_{n,i}^{\top}\Sigma_n^{-2}u_{n,i} - \frac{2}{p}\sum_{i=1}^{p}\widetilde{d}_{n,i}\cdot u_{n,i}^{\top}\Sigma_n^{-1}u_{n,i} + 1\\
&= \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p}\left[\frac{\widetilde{d}_{n,i}^2}{\tau_{n,j}^2} - 2\frac{\widetilde{d}_{n,i}}{\tau_{n,j}} + 1\right]\cdot\left(u_i^{\top}v_j\right)^2\\
&= \int_a^b\int_{-\infty}^{+\infty}\left[\frac{\widetilde{\varphi}_n(x)^2}{t^2} - 2\frac{\widetilde{\varphi}_n(x)}{t} + 1\right]d^2\Theta_n(x,t) ,
\end{aligned}$$

where Θ_n is the random bivariate function from (6)). By applying the technique from the proof of Theorem 3.1 of [22], and by using Theorem 2 to handle the function Θ_n, it follows that

$$\mathcal{L}_n^{Q}(\widetilde{S}_n, \Sigma_n) \xrightarrow{a.s.} \int_a^b\int_{-\infty}^{+\infty}\left[\frac{\widetilde{\varphi}(x)^2}{t^2} - 2\frac{\widetilde{\varphi}(x)}{t} + 1\right]\theta(x,t)\,dx\,dt ,$$

where, as per (7),

$$\forall x \in [a,b] \quad \forall t \in \mathbb{R} \qquad \theta(x,t) := \frac{cxt}{\left|t\left[1-c-cx\breve{m}_F(x)\right]-x\right|^2}\,.$$

Proposition 2 allows us to characterize the asymptotically optimal nonlinear shrinkage function under Quadratic loss.

Corollary 1. *Suppose Assumptions 2–5 hold. A covariance matrix estimator $\widetilde{S}_n$ minimizes in the class of rotation-equivariant estimators the a.s. limit (12) of the Quadratic loss if and only if its limiting shrinkage function $\widetilde{\varphi}$ verifies $\forall x \in Supp(F)$, $\widetilde{\varphi}(x) = \widehat{\varphi}^{Q}(x)$, where*

$$\forall x \in Supp(F) \qquad \widehat{\varphi}^{Q}(x) := \frac{\displaystyle\int_{-\infty}^{+\infty} \frac{1}{t}\cdot\theta(x,t)\,dH(t)}{\displaystyle\int_{-\infty}^{+\infty} \frac{1}{t^2}\cdot\theta(x,t)\,dH(t)}\,. \tag{13}$$

Proof. If we fix $x \in \mathrm{Supp}(F)$, then the marginal contribution of $\widetilde{\varphi}(x)$ to the almost sure (nonrandom) limit of the loss function $\mathcal{L}_n^{Q}(\Sigma_n, \widetilde{S}_n)$ is

$$\int_{-\infty}^{+\infty}\left[\frac{\widetilde{\varphi}(x)^2}{t^2} - 2\frac{\widetilde{\varphi}(x)}{t} + 1\right]\theta(x,t)\,dH(t)\,. \tag{14}$$

The partial derivative of (14) with respect to $\widetilde{\varphi}(x)$ is

$$\int_{-\infty}^{+\infty}\left[\frac{2\widetilde{\varphi}(x)}{t^2} - \frac{2}{t}\right]\theta(x,t)\,dH(t)\,.$$

The first-order condition is

$$\varphi(x)\int_{-\infty}^{+\infty}\frac{1}{t^2}\,\theta(x,t)\,dH(t) = \int_{-\infty}^{+\infty}\frac{1}{t}\,\theta(x,t)\,dH(t)\,.$$

The solution is

$$\varphi(x) = \frac{\displaystyle\int_{-\infty}^{+\infty}\frac{1}{t}\,\theta(x,t)\,dH(t)}{\displaystyle\int_{-\infty}^{+\infty}\frac{1}{t^2}\,\theta(x,t)\,dH(t)}\,.$$

The proof of Theorem 5 is concluded as follows: To the unobservable quantity c corresponds the plug-in estimator p/n; to the unobservable quantity $H(t)$ corresponds the plug-in estimator $\widehat{H}_n(t) := \sum_{i=j}^{p}\mathbb{1}_{[\widehat{\tau}_{n,j},+\infty]}(t)/p$; and to the unobservable quantity $\theta(x)$ corresponds the plug-in estimator $\widehat{\theta}_n(x,t)$ from (8). The fact that these three unobservable quantities can be replaced with their respective plug-in counterparts at no loss asymptotically is established in the same way as in the proof of Theorem 5.2 of [22].

Theorem 6 (*Inverse Quadratic*)**.** *The Inverse Quadratic loss function, which is defined as $\mathcal{L}^{QINV}\big(\widetilde{S}_n, \Sigma_n\big) := \big\|\widetilde{S}_n^{-1}\Sigma_n - \mathbb{I}_n\big\|_F^2/p$, converges almost surely to a deterministic limit that is minimized if $\widetilde{\varphi}_n(\lambda_{n,i})$ is equal to*

$$\widehat{\varphi}_n^{QINV}(\lambda_{n,i}) := \frac{\displaystyle\frac{1}{p}\sum_{j=1}^{p}\widehat{\tau}_{n,j}^2\cdot\widehat{\theta}_n(\lambda_{n,i},\widehat{\tau}_{n,j})}{\displaystyle\frac{1}{p}\sum_{j=1}^{p}\widehat{\tau}_{n,j}\cdot\widehat{\theta}_n(\lambda_{n,i},\widehat{\tau}_{n,j})}\,. \tag{15}$$

The resulting covariance matrix estimator is $\widehat{S}_n^{QINV} := \sum_{i=1}^{p}\widehat{\varphi}_n^{QINV}\big(\lambda_{n,i}\big)\cdot u_{n,i}u_{n,i}^{\top}$.

Proof. To prove the theorem, we need to first state and prove some auxiliary results.

Proposition 3. *Under Assumptions 2–5,*

$$\mathcal{L}_n^{QINV}(\widetilde{S}_n, \Sigma_n) \xrightarrow{\text{a.s.}} \sum_{k=1}^{\kappa}\int_{a_k}^{b_k}\int_{-\infty}^{+\infty}\left[\frac{t^2}{\widetilde{\varphi}(x)^2} - 2\frac{t}{\widetilde{\varphi}(x)} + 1\right]\cdot\theta(x,t)\,dH(t)\,dF(x)\,. \tag{16}$$

Proof. As before, we assume that the support of F is a single compact interval $[a, b] \subset (0, +\infty)$. From Supplementary material B.12 we have:

$$\begin{aligned}\mathcal{L}_n^{\text{QINV}}(\widetilde{S}_n, \Sigma_n) &= \frac{1}{p}\sum_{i=1}^{p} \widetilde{d}_{n,i}^{-2} \cdot u_{n,i}^\top \Sigma_n^2 u_{n,i} - \frac{2}{p}\sum_{i=1}^{p} \widetilde{d}_{n,i}^{-1} \cdot u_{n,i}^\top \Sigma_n u_{n,i} + 1 \\ &= \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p}\left[\frac{\tau_{n,j}^2}{\widetilde{d}_{n,i}^2} - 2\frac{\tau_{n,j}}{\widetilde{d}_{n,i}} + 1\right] \cdot \left(u_i^\top v_j\right)^2 \\ &= \int_a^b \int_{-\infty}^{+\infty} \left[\frac{\tau_{n,j}^2}{\widetilde{\varphi}_n(\lambda_{n,i})^2} - 2\frac{\tau_{n,j}}{\widetilde{\varphi}_n(\lambda_{n,i})} + 1\right] d^2\Theta_n(x,t)\,.\end{aligned}$$

By applying the technique from the proof of Theorem 3.1 of [22], and by using Theorem 2:

$$\mathcal{L}_n^{\text{QINV}}(\widetilde{S}_n, \Sigma_n) \xrightarrow{a.s.} \int_a^b \int_{-\infty}^{+\infty} \left[\frac{t^2}{\widetilde{\varphi}(x)^2} - 2\frac{t}{\widetilde{\varphi}(x)} + 1\right] \theta(x,t)\,dx\,dt\,.$$

Corollary 2. *Under Assumptions 2–5, a covariance matrix estimator $\widetilde{S}_n$ minimizes in the class of rotation-equivariant estimators the a.s. limit (16) of the Inverse Quadratic loss if and only if its limiting shrinkage function $\widetilde{\varphi}$ verifies $\forall x \in Supp(F)$, $\widetilde{\varphi}(x) = \widehat{\varphi}^{QINV}(x)$, where*

$$\forall x \in Supp(F) \qquad \widehat{\varphi}^{QINV}(x) := \frac{\displaystyle\int_{-\infty}^{+\infty} t^2 \cdot \theta(x,t)\,dH(t)}{\displaystyle\int_{-\infty}^{+\infty} t \cdot \theta(x,t)\,dH(t)}\,.$$

Proof. If we fix $x \in \text{Supp}(F)$, then the marginal contribution of $\widetilde{\varphi}(x)$ to the almost sure (nonrandom) limit of the loss function $\mathcal{L}_n^{\text{QINV}}(\Sigma_n, \widetilde{S}_n)$ is

$$\int_{-\infty}^{+\infty} \left[\frac{t^2}{\widetilde{\varphi}(x)^2} - 2\frac{t}{\widetilde{\varphi}(x)} + 1\right] \theta(x,t)\,dH(t)\,. \tag{17}$$

The partial derivative of (17) with respect to $\widetilde{\varphi}(x)$ is

$$\int_{-\infty}^{+\infty} \left[-2\frac{t^2}{\widetilde{\varphi}(x)^3} + 2\frac{t}{\widetilde{\varphi}(x)^2}\right] \theta(x,t)\,dH(t)\,.$$

The first-order condition is

$$\varphi(x)\int_{-\infty}^{+\infty} t\,\theta(x,t)\,dH(t) = \int_{-\infty}^{+\infty} t^2\,\theta(x,t)\,dH(t)\,.$$

The solution is

$$\varphi(x) = \frac{\displaystyle\int_{-\infty}^{+\infty} t^2\,\theta(x,t)\,dH(t)}{\displaystyle\int_{-\infty}^{+\infty} t\,\theta(x,t)\,dH(t)}\,.$$

The proof of Theorem 6 is concluded as before: To the unobservable quantity c corresponds the plug-in estimator p/n; to the unobservable quantity $H(t)$ corresponds the plug-in estimator $\widehat{H}_n(t) := \sum_{i=j}^{p} \mathbb{1}_{[\widehat{\tau}_{n,j}, +\infty]}(t)/p$; and to the unobservable quantity $\theta(x)$ corresponds the plug-in estimator $\widehat{\theta}_n(x,t)$ from (8). The fact that these three unobservable quantities can be replaced with their respective plug-in counterparts at no loss asymptotically is established in the same way as in the proof of Theorem 5.2 of [22].

4.2. Two infinite families of loss functions

We have so far covered 12 loss functions, including many of the classic ones, from which we have derived a total of 7 different optimal nonlinear shrinkage formulas (as there are some commonalities). It is tedious to keep adding more by hand. Most applied researchers should have already been able to find 'the shoe that fits' in this rather extensive list by now.

If not, the only systematic method is to study an (uncountably) infinite number of loss functions, and to find the nonlinear shrinkage formula exactly optimized with respect to each of them. To the best of our knowledge, an ambitious project on this scale has never been envisioned before. In doing so, we will meet again some old acquaintances: 6 of

the 12 loss functions already analyzed manually are special cases of the two general theorems presented below. The first infinite family of loss functions is what we call Generalized Frobenius.

Theorem 7 (*Generalized Frobenius*). *For any invertible and continuously differentiable function γ defined on $(0, +\infty)$, the Generalized Frobenius loss $\mathcal{L}_n^{\gamma,F}(\widetilde{S}_n, \Sigma_n) := \left\|\gamma\left(\widetilde{S}_n\right) - \gamma\left(\Sigma_n\right)\right\|_F^2 / p$ converges almost surely to a deterministic limit that is minimized if $\widetilde{\varphi}_n(\lambda_{n,i})$ is equal to*

$$\widehat{\varphi}_n^{\gamma}(\lambda_{n,i}) := \gamma^{-1}\left(\frac{1}{p}\sum_{j=1}^{p}\gamma\left(\widehat{\tau}_{n,j}\right)\cdot\widehat{\theta}_n(\lambda_{n,i},\widehat{\tau}_{n,j})\right) . \tag{18}$$

The resulting covariance matrix estimator is $\widehat{S}_n^{\gamma} := \sum_{i=1}^{p}\widehat{\varphi}_n^{\gamma}\left(\lambda_{n,i}\right)\cdot u_{n,i}u_{n,i}^{\top}$.

Proof. To prove the theorem, we need to first state and prove some auxiliary results.

Proposition 4. *Under Assumptions 2–5,*

$$\begin{aligned}\mathcal{L}_n^{\gamma,F}(\widetilde{S}_n, \Sigma_n) \xrightarrow{\text{a.s.}} &\int_{-\infty}^{+\infty}\gamma(t)^2\,dH(t) - 2\sum_{k=1}^{\kappa}\int_{a_k}^{b_k}\int_{-\infty}^{+\infty}\gamma(t)\,\gamma\left(\widetilde{\varphi}(x)\right)\cdot\theta(x,t)\,dH(t)\,dF(x)\\ &+ \sum_{k=1}^{\kappa}\int_{a_k}^{b_k}\gamma\left(\widetilde{\varphi}(x)\right)^2 dF(x) .\end{aligned} \tag{19}$$

Proof. For simplicity: $\text{Supp}(F) = [a, b] \subset (0, +\infty)$. From Supplementary material B.13 we have:

$$\begin{aligned}\mathcal{L}_n^{\gamma,F}(\widetilde{S}_n, \Sigma_n) &= \frac{1}{p}\sum_{i=1}^{p}\left\{\gamma(\tau_{n,i})^2 - 2u_{n,i}^{\top}\gamma(\Sigma_n)u_{n,i}\gamma(\widetilde{d}_{n,i}) + \gamma(\widetilde{d}_{n,i})^2\right\}\\ &= \frac{1}{p}\sum_{j=1}^{p}\gamma(\tau_{n,j})^2 - \frac{2}{p}\sum_{i=1}^{p}\sum_{j=1}^{p}\gamma\left(\tau_{n,j}\right)\gamma\left(\widetilde{\varphi}(\lambda_{n,i})\right)\cdot\left(u_{n,i}^{\top}v_{n,j}\right)^2 + \frac{1}{p}\sum_{i=1}^{p}\gamma\left(\widetilde{\varphi}(\lambda_{n,i})\right)^2\\ &= \int_{-\infty}^{+\infty}\gamma(t)^2 dH_n(t) - 2\int_a^b\int_{-\infty}^{+\infty}\gamma(t)\gamma\left(\widetilde{\varphi}_n(x)\right)d^2\Theta_n(x,t) + \int_a^b\gamma\left(\widetilde{\varphi}_n(x)\right)^2 dF_n(x) .\end{aligned}$$

By applying the technique from the proof of Theorem 3.1 of [22], and by using Theorem 2:

$$\begin{aligned}\mathcal{L}_n^{\gamma,F}(\widetilde{S}_n, \Sigma_n) \xrightarrow{\text{a.s.}} &\int_{-\infty}^{+\infty}\gamma(t)^2\,dH(t) - 2\sum_{k=1}^{\kappa}\int_{a_k}^{b_k}\int_{-\infty}^{+\infty}\gamma(t)\,\gamma\left(\widetilde{\varphi}(x)\right)\cdot\theta(x,t)\,dH(t)\,dF(x)\\ &+ \sum_{k=1}^{\kappa}\int_{a_k}^{b_k}\gamma\left(\widetilde{\varphi}(x)\right)^2 dF(x) .\end{aligned}$$

Corollary 3. *Suppose Assumptions 2–5 hold. A covariance matrix estimator $\widetilde{S}_n$ minimizes in the class of rotation-equivariant estimators the a.s. limit (19) of the Generalized Frobenius loss if and only if its limiting shrinkage function $\widetilde{\varphi}$ verifies $\forall x \in \text{Supp}(F)$, $\widetilde{\varphi}(x) = \varphi^{\gamma}(x)$, where*

$$\forall x \in \text{Supp}(F) \qquad \varphi^{\gamma}(x) := \gamma^{-1}\left[\int_{-\infty}^{+\infty}\gamma(t)\cdot\theta(x,t)dH(t)\right]. \tag{20}$$

This yields an oracle covariance matrix estimator $S_n^{\gamma} := U_n\text{Diag}\left(\varphi^{\gamma}(\lambda_{n,1}), \ldots, \varphi^{\gamma}(\lambda_{n,p})\right)U_n^{\top}$.

Proof. If we fix $x \in \text{Supp}(F)$, then the marginal contribution of $\widetilde{\varphi}(x)$ to the almost sure (nonrandom) limit of the loss function $\mathcal{L}_n^{\gamma,F}(\Sigma_n, \widetilde{S}_n)$is

$$-2\int_{-\infty}^{+\infty}\gamma(t)\gamma\left(\widetilde{\varphi}(x)\right)\cdot\theta(x,t)\,dH(t) + \gamma\left(\widetilde{\varphi}(x)\right)^2 . \tag{21}$$

The partial derivative of (21) with respect to $\widetilde{\varphi}(x)$ is

$$-2\int_{-\infty}^{+\infty}\gamma(t)\gamma^{\top}\left(\widetilde{\varphi}(x)\right)\cdot\theta(x,t)\,dH(t) + 2\gamma^{\top}\left(\widetilde{\varphi}(x)\right)\gamma\left(\widetilde{\varphi}(x)\right) .$$

The first-order condition is $\gamma\big(\widetilde{\varphi}(x)\big) = \int_{-\infty}^{+\infty} \gamma(t) \cdot \theta(x,t)\, dH(t)$, hence the solution is

$$\widetilde{\varphi}(x) = \gamma^{-1}\left(\int_{-\infty}^{+\infty} \gamma(t) \cdot \theta(x,t)\, dH(t)\right) .$$

The proof of Theorem 7 is concluded as before: To the unobservable quantity c corresponds the plug-in estimator p/n; to the unobservable quantity $H(t)$ corresponds the plug-in estimator $\widehat{H}_n(t) := \sum_{i=j}^{p} \mathbb{1}_{[\widehat{\tau}_{n,j},+\infty]}(t)/p$; and to the unobservable quantity $\theta(x)$ corresponds the plug-in estimator $\widehat{\theta}_n(x,t)$ from (8). The fact that these three unobservable quantities can be replaced with their respective plug-in counterparts at no loss asymptotically is established in the same way as in the proof of Theorem 5.2 of [22].

Remark 4. The Frobenius, Inverse Frobenius, Log-Euclidian, and Fréchet losses are special cases of the General Frobenius family, corresponding, respectively, to $\gamma(x)$ equal to x, $1/x$, $\log(x)$, and $\sqrt{x}$.

A second infinite family of loss functions is based on the Kullback–Leibler divergence of [16]. Given two multivariate normal distributions $\mathcal{N}(0, A_i)$ with zero mean and covariance matrix A_i, for $i \in \{1,2\}$, their dimension-normalized Kullback–Leibler divergence is:

$$D_{KL}\big(\mathcal{N}(0,A_1)\|\,\mathcal{N}(0,A_2)\big) := \frac{1}{2p}\left\{\mathrm{Tr}\big[A_2^{-1}A_1\big] - \log\big[\det(A_2^{-1}A_1)\big] - p\right\} .$$

Stein's loss and the Inverse Stein loss are special cases of the Generalized Kullback–Leibler family defined below, obtained by setting $\gamma(x)$ equal to $1/x$ and x, respectively.

Theorem 8 (*Generalized Kullback–Leibler*)**.** *For any invertible and continuously differentiable function γ defined on $(0,+\infty)$, the Generalized Kullback–Leibler loss function*

$$\mathcal{L}_n^{\gamma,KL}(\widetilde{S}_n, \Sigma_n) := \frac{1}{2p}\left\{Tr\left[\gamma\big(\widetilde{S}_n\big)^{-1}\gamma\big(\Sigma_n\big)\right] - \log\det\left[\gamma(\widetilde{S}_n)^{-1}\gamma(\Sigma_n)\right] - p\right\}$$

converges almost surely to a deterministic limit that is minimized if $\widetilde{\varphi}_n(\lambda_{n,i})$ is equal to the quantity $\widehat{\varphi}_n^{\gamma}(\lambda_{n,i})$ defined in (18), $i \in \{1,\ldots,p\}$.

Proof. To prove the theorem, we need to first state and prove some auxiliary results.

Proposition 5. *Under Assumptions 2–5,*

$$\begin{aligned}\mathcal{L}_n^{\gamma,KL}(\widetilde{S}_n, \Sigma_n) \xrightarrow{\text{a.s.}} & \frac{1}{2}\sum_{k=1}^{\kappa}\int_{a_k}^{b_k}\int_{-\infty}^{+\infty}\frac{\gamma(t)}{\gamma\big(\widetilde{\varphi}(x)\big)}\cdot\theta(x,t)\,dH(t)\,dF(x) \\ & + \frac{1}{2}\sum_{k=1}^{\kappa}\int_{a_k}^{b_k}\log\big[\gamma\big(\widetilde{\varphi}(x)\big)\big]\,dF(x) - \frac{1}{2}\int_{-\infty}^{+\infty}\log\big[\gamma(t)\big]\,dH(t) - \frac{1}{2} . \end{aligned} \tag{22}$$

Proof. For simplicity, $\mathrm{Supp}(F) = [a,b] \subset (0,+\infty)$. From Supplementary material B.14:

$$\begin{aligned}\mathcal{L}_n^{\gamma,\mathrm{KL}}(\widetilde{S}_n, \Sigma_n) &= \frac{1}{2p}\sum_{i=1}^{p}\left\{\frac{u_i^{\top}\gamma(\Sigma)u_i}{\gamma\big(\widetilde{\delta}_i\big)} + \log\big[\gamma\big(\widetilde{\delta}_i\big)\big] - \log\big[\gamma\big(\tau_i\big)\big] - 1\right\} \\ &= \frac{1}{2}\int_a^b\int_{-\infty}^{+\infty}\frac{\gamma(t)}{\gamma\big(\widetilde{\varphi}_n(x)\big)}\,d^2\Theta_n(x,t) + \frac{1}{2}\int_a^b\log\big[\gamma\big(\widetilde{\varphi}_n(x)\big)\big]\,dF_n(x) \\ &\quad - \frac{1}{2}\int_{-\infty}^{+\infty}\log\big[\gamma(t)\big]dH_n(t) - \frac{1}{2} .\end{aligned}$$

By applying the technique from the proof of Theorem 3.1 of [22], and by using Theorem 2:

$$\begin{aligned}\mathcal{L}_n^{\gamma,\mathrm{KL}}(\widetilde{S}_n, \Sigma_n) \xrightarrow{\text{a.s.}} & \frac{1}{2}\sum_{k=1}^{\kappa}\int_{a_k}^{b_k}\int_{-\infty}^{+\infty}\frac{\gamma(t)}{\gamma\big(\widetilde{\varphi}(x)\big)}\cdot\theta(x,t)\,dH(t)\,dF(x) \\ & + \frac{1}{2}\sum_{k=1}^{\kappa}\int_{a_k}^{b_k}\log\big[\gamma\big(\widetilde{\varphi}(x)\big)\big]\,dF(x) - \frac{1}{2}\int_{-\infty}^{+\infty}\log\big[\gamma(t)\big]\,dH(t) - \frac{1}{2} .\end{aligned}$$

Corollary 4. *Suppose Assumptions 2–5 hold. A covariance matrix estimator $\widetilde{S}_n$ minimizes in the class of rotation-equivariant estimators the a.s. limit (22) of the Generalized Kullback–Leibler loss if and only if its limiting shrinkage function $\widetilde{\varphi}$ verifies*

$\forall x \in Supp(F)$, $\widetilde{\varphi}(x) = \varphi^{\gamma}(x)$, *where* $\varphi^{\gamma}(x)$*is defined by* (20)*. This results in the same oracle covariance matrix estimator* $S_n^{\gamma} := U_n Diag\big(\varphi^{\gamma}(\lambda_{n,1}), \ldots, \varphi^{\gamma}(\lambda_{n,p})\big)U_n^{\top}$ *as in Corollary* 3*.*

Proof. If we fix $x \in \text{Supp}(F)$, then the marginal contribution of $\widetilde{\varphi}(x)$ to the almost sure (nonrandom) limit of the loss function $\mathcal{L}_n^{\gamma,\text{KL}}(\Sigma_n, \widetilde{S}_n)$is

$$\frac{1}{2}\int_{-\infty}^{+\infty} \frac{\gamma(t)}{\gamma\big(\widetilde{\varphi}(x)\big)} \cdot \theta(x,t)\, dH(t) + \log\big[\gamma\big(\widetilde{\varphi}(x)\big)\big] \ . \tag{23}$$

The partial derivative of (23) with respect to $\widetilde{\varphi}(x)$ is

$$-\frac{1}{2}\int_{-\infty}^{+\infty} \frac{\gamma(t)}{\gamma\big(\widetilde{\varphi}(x)\big)^2}\gamma^{\top}\big(\widetilde{\varphi}(x)\big) \cdot \theta(x,t)\, dH(t) + \frac{\gamma^{\top}\big(\widetilde{\varphi}(x)\big)}{2\gamma\big(\widetilde{\varphi}(x)\big)} \ .$$

The first-order condition is $\gamma\big(\widetilde{\varphi}(x)\big) = \int_{-\infty}^{+\infty} \gamma(t) \cdot \theta(x,t)\, dH(t)$, hence the solution is

$$\widetilde{\varphi}(x) = \gamma^{-1}\left(\int_{-\infty}^{+\infty} \gamma(t) \cdot \theta(x,t)\, dH(t)\right) \ .$$

The proof of Theorem 8 is concluded as before, by showing that replacing the key oracle objects with their plug-in counterparts comes at no cost under large-dimensional asymptotics.

5. Singular case: $p > n$

This is a case of great practical importance. When it happens, the sample covariance matrix is singular: It has $p - n$ eigenvalues equal to zero and is thus only positive semi-definite. There then exist some linear combinations of the original variables that falsely appear to have zero variance when one only looks in-sample. In a sense, the sample covariance matrix, with its $p(p+1)/2$ degrees of freedom, 'overfits' the data set of size $n \times p$.

5.1. Analysis in finite samples

With respect to the loss functions studied in this paper, the optimal nonlinear shrinkage formula applied to the n nonzero sample eigenvalues remains the same as in the case $p < n$, so no need to revisit. The only item to be determined is how to shrink the $p - n$ null sample eigenvalues. Recall that we sort the sample eigenvalues in nondecreasing order without loss of generality (w.l.o.g.,) so the null eigenvalues are the first $p - n$ ones. To build intuition, we start as before with the finite-sample case: Table 4 presents a counterpart to Tables 1–3, listing how to optimally shrink the null sample eigenvalues.

The pattern is clear: compute how the eigenvectors in the null space of the sample covariance matrix relate to (a function of) the population covariance matrix, take the average(s), and take smooth transformations of the average(s), where a transformation could be simply the identity. There is a rotational indeterminacy in this null space of dimension $p - n$, but the formulas in the last column are invariant to a rotation of the basis of the null eigenvectors, so it does not matter.

5.2. Analysis under large-dimensional asymptotics

Assumption 6 (*Singular*)**.** The ratio p/n converges, as $n \to \infty$, to a finite limit $c > 1$. Furthermore, there exists a compact interval included in $(1, +\infty)$ that contains p/n for all n large enough.

Given that the first $p - n$ sample eigenvalues are devoid of informational content, it is judicious to focus on the e.d.f of the n other ones: $\forall x \in \mathbb{R}$ $\underline{F}_n(x) := \frac{1}{n}\sum_{i=p-n+1}^{p} \mathbb{1}_{[\lambda_{n,i},+\infty)}(x)$. Under Assumptions 3–6, it admits a nonrandom limit:

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

Of particular interest will be its Stieltjes transform: $\forall z \in \mathbb{C}^+$ $m_{\underline{F}}(z) := \int \frac{1}{\lambda - z}\, d\underline{F}(\lambda)$, which admits a continuous extension onto the real line: $\forall x \in \mathbb{R}$ $\breve{m}_{\underline{F}}(x) := \lim_{z \in \mathbb{C}^+ \to x} m_{\underline{F}}(z)$.

5.3. Optimal shrinkage of null sample eigenvalues

At this stage, what we need is an equivalent of (2) that pertains to the shrinkage of the null sample eigenvalues. It comes from Theorem 9 of [17]:

$$\frac{1}{p-n}\sum_{i=1}^{p-n} u_{n,i}^{\top}\gamma(\Sigma_n)u_{n,i} \approx \frac{1}{p}\sum_{j=1}^{p} \gamma(\widehat{\tau}_{n,j}) \cdot \frac{1}{\left(1 - \frac{n}{p}\right)\left[1 + \underline{\breve{m}}_{n,p}^{\widehat{\tau}_n}(0)\widehat{\tau}_{n,j}\right]} \ , \tag{24}$$

Table 4
Formulas for shrinking null sample eigenvalues. The first column gives the name of the loss function; the second column gives the corresponding stylized formula; and the third column gives the shrinkage formula for $\widetilde{d}_i$, that is, the replacement for the ith sample eigenvalue. This shrinkage formula only applies to the null sample eigenvalues, that is, to the first $p-n$ sample eigenvalues.

Loss function	Stylized Formula	Null Shrinkage
Frobenius	$\left\|\widetilde{S}-\Sigma\right\|_{\mathrm{F}}$	$\frac{1}{p-n}\sum_{i=1}^{p-n} u_i^\top \Sigma u_i$
Inverse Stein	$\mathrm{Tr}\left[\widetilde{S}^{-1}\Sigma\right]-\log\left[\det\left(\widetilde{S}^{-1}\Sigma\right)\right]$	$\frac{1}{p-n}\sum_{i=1}^{p-n} u_i^\top \Sigma u_i$
Minimum Variance	$\mathrm{Tr}\left[\widetilde{S}^{-1}\Sigma\widetilde{S}^{-1}\right]\Big/\left(\mathrm{Tr}\left[\widetilde{S}^{-1}\right]\right)^2$	$\frac{1}{p-n}\sum_{i=1}^{p-n} u_i^\top \Sigma u_i$
Stein	$\mathrm{Tr}\left[\widetilde{S}\Sigma^{-1}\right]-\log\left[\det\left(\widetilde{S}\Sigma^{-1}\right)\right]$	$\left(\frac{1}{p-n}\sum_{i=1}^{p-n} u_i^\top \Sigma^{-1} u_i\right)^{-1}$
Inverse Frobenius	$\left\|\widetilde{S}^{-1}-\Sigma^{-1}\right\|_{\mathrm{F}}$	$\left(\frac{1}{p-n}\sum_{i=1}^{p-n} u_i^\top \Sigma^{-1} u_i\right)^{-1}$
Symmetrized Stein	$\mathrm{Tr}\left[\widetilde{S}\Sigma^{-1}+\widetilde{S}^{-1}\Sigma\right]$	$\sqrt{\frac{\frac{1}{p-n}\sum_{i=1}^{p-n} u_i^\top \Sigma u_i}{\frac{1}{p-n}\sum_{i=1}^{p-n} u_i^\top \Sigma^{-1} u_i}}$
Weighted Frobenius	$\mathrm{Tr}\left[\left(\widetilde{S}-\Sigma\right)^2\Sigma^{-1}\right]$	$\left(\frac{1}{p-n}\sum_{i=1}^{p-n} u_i^\top \Sigma^{-1} u_i\right)^{-1}$
Disutility	$\mathrm{Tr}\left[\left(\widetilde{S}^{-1}-\Sigma^{-1}\right)^2\Sigma\right]$	$\frac{1}{p-n}\sum_{i=1}^{p-n} u_i^\top \Sigma u_i$
Log-Euclidian	$\left\|\log(\widetilde{S})-\log(\Sigma)\right\|_{\mathrm{F}}$	$\exp\left[\frac{1}{p-n}\sum_{i=1}^{p-n} u_i^\top \log(\Sigma) u_i\right]$
Fréchet	$\left\|\widetilde{S}^{1/2}-\Sigma^{1/2}\right\|_{\mathrm{F}}$	$\left(\frac{1}{p-n}\sum_{i=1}^{p-n} u_i^\top \Sigma^{1/2} u_i\right)^2$
Quadratic	$\left\|\Sigma^{-1}\widetilde{S}-\mathbb{I}\right\|_{\mathrm{F}}$	$\frac{\frac{1}{p-n}\sum_{i=1}^{p-n} u_i^\top \Sigma^{-1} u_i}{\frac{1}{p-n}\sum_{i=1}^{p-n} u_i^\top \Sigma^{-2} u_i}$
Inverse Quadratic	$\left\|\widetilde{S}^{-1}\Sigma-\mathbb{I}\right\|_{\mathrm{F}}$	$\frac{\frac{1}{p-n}\sum_{i=1}^{p-n} u_i^\top \Sigma^{2} u_i}{\frac{1}{p-n}\sum_{i=1}^{p-n} u_i^\top \Sigma u_i}$

where $\widehat{\boldsymbol{\tau}}_n := \{\widehat{\tau}_{n,j}\}_{j=1}^p$ is, as before, the estimator of population eigenvalues obtained by numerically inverting the QuEST function, and $\underline{\breve{m}}_{n,p}^{\widehat{\boldsymbol{\tau}}_n}(0)$ is a strongly consistent estimator of $\breve{m}_{\underline{F}}(0)$ which is another by-product of the QuEST function (when $p > n$). As per Section 3.2.2 of [20], $\underline{\breve{m}}_{n,p}^{\widehat{\boldsymbol{\tau}}_n}(0)$ is the unique solution $m \in (0,\infty)$ to the equation

$$\frac{1}{m} = \frac{1}{n}\sum_{j=1}^{p}\frac{\widehat{\tau}_{n,j}}{1+\widehat{\tau}_{n,j}\,m}\ . \tag{25}$$

(24) enables us to extend the squared-dot-product function $\theta(x,t)$ presented in Section 3.3 to handle the case $x=0$. Fig. 2 graphs

$$\theta(0,t) := \frac{1}{\left(1-\frac{1}{c}\right)\left[1+\breve{m}_{\underline{F}}(0)\,t\right]} \tag{26}$$

as a function of t for various values of the concentration ratio p/n. We use the same baseline scenario as in Fig. 1: the population eigenvalues are evenly spread in the interval [1,5].

Eigenvectors in the null space of the sample covariance matrix tend to be more (less) aligned with population eigenvectors corresponding to small (large) population eigenvalues, which makes intuitive sense. The degree of preferential alignment is inversely related to the concentration ratio, as a high ratio p/n disorients the sample eigenvectors. The overall pattern is highly nonlinear, and could only be pinned down through (25)–(26) from RMT. Note that, by construction, the dimension-normalized density of the squared dot-product averages to 1, so it is deviations from the baseline number of 1 that are informative.

5.4. Covariance matrix estimation in the singular case

Theorems 3–8 remain valid when $c > 1$, with the understanding that the estimator of the squared dot-product in the null space of the sample covariance matrix, $i \in \{1,\ldots,p-n\}$, is, for $j \in \{1,\ldots,p\}$,

$$\widehat{\theta}_n(\lambda_{n,i},\widehat{\tau}_{n,j}) = \widehat{\theta}_n(0,\widehat{\tau}_{n,j}) := \frac{1}{\left(1-\frac{n}{p}\right)\left[1+\underline{\breve{m}}_{n,p}^{\widehat{\boldsymbol{\tau}}}(0)\widehat{\tau}_{n,j}\right]}\ . \tag{27}$$

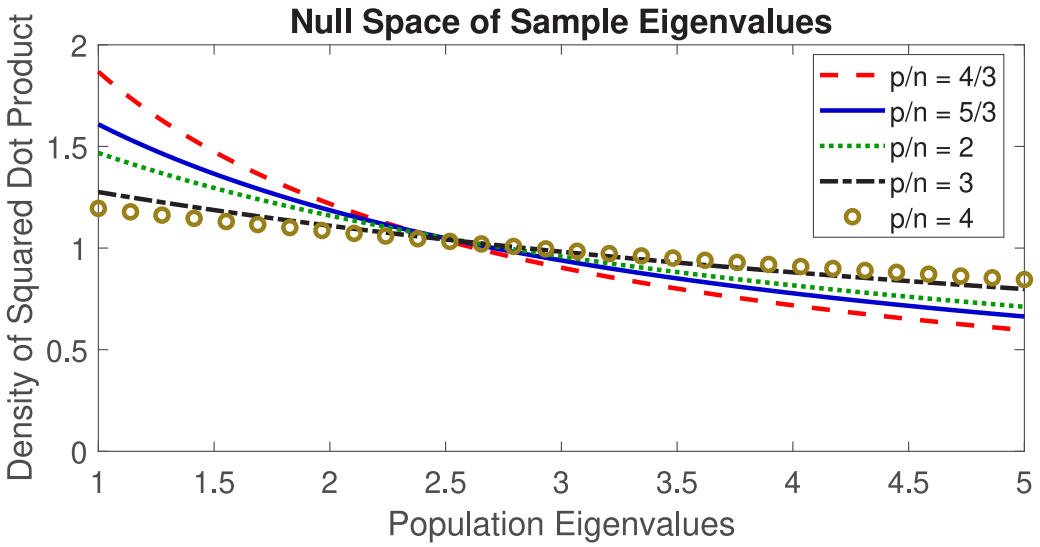

Fig. 2. The Radon–Nikodym derivative $\theta(0,t)$ as a function of the population eigenvalues. This plot shows how aligned the null-space sample eigenvectors are with the population eigenvectors.

In order to show how this works, we need only state and prove the singular-case counterpart of Theorem 7, as the other theorems are adapted from $p < n$ to the $p > n$ case in similar fashion.

Theorem 9. *Under Assumptions 3–6, the Generalized Frobenius loss admits an almost sure (deterministic) limit, which is minimized by the nonlinear shrinkage formula*

$$\widehat{\varphi}_n^{\gamma}(\lambda_{n,i}) := \gamma^{-1}\left(\frac{1}{p}\sum_{j=1}^{p}\gamma\left(\widehat{\tau}_{n,j}\right)\cdot\widehat{\theta}_n(\lambda_{n,i},\widehat{\tau}_{n,j})\right) \ ,$$

where the bivariate function $\widehat{\theta}_n(x,t)$ is given by (8) for $x > 0$, and $\widehat{\theta}_n(0,t)$ is given by (27). The resulting covariance matrix estimator is $\widehat{S}_n^{\gamma} := \sum_{i=1}^{p}\widehat{\varphi}_n^{\gamma}\left(\lambda_{n,i}\right)\cdot u_{n,i}u_{n,i}^{\top}$.

Proof. To prove the theorem, we need to first state and prove some auxiliary results. There is one small difference with the non-singular case: the support of the limiting sample spectral distribution F is now $\mathrm{Supp}(F) = \{0\}\cup\left(\bigcup_{k=1}^{\kappa}[a_k,b_k]\right)$, where (as before) $0 < a_1 < b_1 < \cdots < a_\kappa < b_\kappa < \infty$.

Proposition 6. *Under Assumptions 3–6,*

$$\begin{aligned}\mathcal{L}_n^{\gamma,F}(\widetilde{S}_n,\Sigma_n) \xrightarrow{\text{a.s.}} &\int_{-\infty}^{+\infty}\gamma(t)^2\,dH(t) - 2\sum_{k=1}^{\kappa}\int_{a_k}^{b_k}\int_{-\infty}^{+\infty}\gamma(t)\,\gamma\left(\widetilde{\varphi}(x)\right)\cdot\theta(x,t)\,dH(t)\,dF(x)\\ &- 2\frac{c-1}{c}\int_{-\infty}^{+\infty}\gamma(t)\,\gamma\left(\widetilde{\varphi}(0)\right)\cdot\theta(0,t)\,dH(t) + \sum_{k=1}^{\kappa}\int_{a_k}^{b_k}\gamma\left(\widetilde{\varphi}(x)\right)^2\,dF(x) + \frac{c-1}{c}\gamma\left(\widetilde{\varphi}(0)\right)^2 \ , \end{aligned} \tag{28}$$

where $\theta(0,t)$ is given by (26).

Proof. For simplicity, we assume that $\mathrm{Supp}(F) = \{0\}\cup[a,b]$, as the extension to the case $\kappa > 1$ is straightforward. Starting from the proof of Proposition 4:

$$\begin{aligned}\mathcal{L}_n^{\gamma,\mathrm{F}}(\Sigma_n,\widetilde{S}_n) &= \frac{1}{p}\sum_{j=1}^{p}\gamma(\tau_{n,j})^2 - \frac{2}{p}\sum_{i=1}^{p}\sum_{j=1}^{p}\gamma\left(\tau_{n,j}\right)\gamma\left(\widetilde{\varphi}_n(\lambda_{n,i})\right)\cdot\left(u_{n,i}^{\top}v_{n,j}\right)^2 + \frac{1}{p}\sum_{i=1}^{p}\gamma\left(\widetilde{\varphi}_n(\lambda_{n,i})\right)^2\\ &= \frac{1}{p}\sum_{j=1}^{p}\gamma(\tau_{n,j})^2 - \frac{2}{p}\sum_{i=p-n+1}^{p}\sum_{j=1}^{p}\gamma\left(\tau_{n,j}\right)\gamma\left(\widetilde{\varphi}_n(\lambda_{n,i})\right)\cdot\left(u_{n,i}^{\top}v_{n,j}\right)^2\end{aligned}$$

$$
\begin{aligned}
&- \frac{2}{p}\sum_{i=1}^{p-n}\sum_{j=1}^{p}\gamma\left(\tau_{n,j}\right)\gamma\left(\widetilde{\varphi}_n(0)\right)\cdot\left(u_{n,i}^{\top}v_{n,j}\right)^2+\frac{1}{p}\sum_{i=p-n+1}^{p}\gamma\left(\widetilde{\varphi}_n(\lambda_{n,i})\right)^2+\frac{1}{p}\sum_{i=1}^{p-n}\gamma\left(\widetilde{\varphi}_n(0)\right)^2\\
&=\int_{-\infty}^{+\infty}\gamma(t)^2dH_n(t)-2\int_a^b\int_{-\infty}^{+\infty}\gamma(t)\gamma\left(\widetilde{\varphi}_n(x)\right)d^2\Theta_n(x,t)\\
&\quad-2\frac{p-n}{p}\int_{-\infty}^{+\infty}\gamma(t)\gamma\left(\widetilde{\varphi}_n(0)\right)d\Theta_n(0,t)+\int_a^b\gamma\left(\widetilde{\varphi}_n(x)\right)^2dF_n(x)+\frac{p}{p-n}\gamma\left(\widetilde{\varphi}_n(0)\right)^2.
\end{aligned}
$$

By applying the technique from the proof of Theorem 6.1 of [22], and by using Theorem 3 of [17] to handle the limit of Θ_n, it follows that:

$$
\begin{aligned}
\mathcal{L}_n^{\gamma,\mathrm{F}}(\Sigma_n,\widetilde{S}_n)\xrightarrow{\text{a.s.}}&\int_{-\infty}^{+\infty}\gamma(t)^2\,dH(t)-2\sum_{k=1}^{\kappa}\int_{a_k}^{b_k}\int_{-\infty}^{+\infty}\gamma(t)\,\gamma\left(\widetilde{\varphi}(x)\right)\cdot\theta(x,t)\,dH(t)\,dF(x)\\
&-2\frac{c-1}{c}\int_{-\infty}^{+\infty}\gamma(t)\,\gamma\left(\widetilde{\varphi}(0)\right)\cdot\frac{1}{\left(1-\frac{1}{c}\right)\left[1+\breve{m}_{\underline{F}}(0)\,t\right]}\,dH(t)\\
&+\sum_{k=1}^{\kappa}\int_{a_k}^{b_k}\gamma\left(\widetilde{\varphi}(x)\right)^2dF(x)+\frac{c-1}{c}\gamma\left(\widetilde{\varphi}(0)\right)^2.
\end{aligned}
\tag{29}
$$

Corollary 5. *Under Assumptions 3–6, a covariance matrix estimator $\widetilde{S}_n$ minimizes in the class of rotation-equivariant estimators the a.s. limit (28) of the Generalized Frobenius loss if and only if its limiting shrinkage function $\widetilde{\varphi}$ verifies $\forall x\in Supp(F)$, $\widetilde{\varphi}(x)=\varphi^{\gamma}(x)$, where*

$$
\forall x\in Supp(F)\qquad \varphi^{\gamma}(x):=\gamma^{-1}\left[\int_{-\infty}^{+\infty}\gamma(t)\cdot\theta(x,t)dH(t)\right].
\tag{30}
$$

This yields the oracle covariance matrix estimator $S_n^{\gamma}:=U_nDiag\left(\varphi^{\gamma}(\lambda_{n,1}),\ldots,\varphi^{\gamma}(\lambda_{n,p})\right)U_n^{\top}$.

For $x\in\bigcup_{k=1}^{\kappa}[a_k,b_k]$, the proof is the same as for Corollary 3. The only pending matter is what happens when $x=0$. The marginal contribution of $\widetilde{\varphi}(0)$ to the almost sure (nonrandom) limit of the loss function $\mathcal{L}_n^{\gamma,\mathrm{F}}(\Sigma_n,\widetilde{S}_n)$ is

$$
-2\frac{c-1}{c}\int_{-\infty}^{+\infty}\gamma(t)\gamma\left(\widetilde{\varphi}(0)\right)\cdot\theta(0,t)\,dH(t)+\frac{c-1}{c}\gamma\left(\widetilde{\varphi}(0)\right)^2.
\tag{31}
$$

The partial derivative of (31) with respect to $\widetilde{\varphi}(0)$ is

$$
-2\frac{c-1}{c}\int_{-\infty}^{+\infty}\gamma(t)\gamma^{\top}\left(\widetilde{\varphi}(0)\right)\cdot\theta(0,t)\,dH(t)+2\frac{c-1}{c}\gamma^{\top}\left(\widetilde{\varphi}(0)\right)\gamma\left(\widetilde{\varphi}(0)\right).
$$

The first-order condition is $\gamma\left(\widetilde{\varphi}(0)\right)=\int_{-\infty}^{+\infty}\gamma(t)\cdot\theta(0,t)\,dH(t)$, hence the solution is

$$
\widetilde{\varphi}(0)=\gamma^{-1}\left(\int_{-\infty}^{+\infty}\gamma(t)\cdot\theta(0,t)\,dH(t)\right).
$$

The proof of Theorem 9 is concluded as follows: To the unobservable quantity c corresponds the plug-in estimator p/n; to the unobservable function $H(t)$ corresponds the plug-in estimator $\widehat{H}_n(t):=\sum_{i=j}^{p}\mathbb{1}_{[\widehat{\tau}_{n,j},+\infty]}(t)/p$; to the unobservable function $\theta(x,t)$ corresponds the plug-in estimator $\widehat{\theta}_n(x,t)$ from (8) for $x>0$; and to the unobservable quantity $\theta(0,t)$ corresponds the plug-in estimator $\widehat{\theta}_n(0,t)$ from (27). The fact that these four unobservables can be replaced with their respective plug-in counterparts at no loss asymptotically is established in the same way as in the proof of Theorem 6.2 of [22].

Remark 5. Both infinite families of loss functions confirm the asymptotic optimality of the same infinite family of nonlinear shrinkage estimators $\widehat{S}_n^{\gamma}$. The Frobenius norm is important because it is just the Euclidian distance on the space of matrices, whereas the Kullback–Leibler divergence is important in a completely different field: information theory. Two justifications coming from such different perspectives combine to give strong backing to the covariance matrix estimator $\widehat{S}_n^{\gamma}$, no matter which function γ the end-user is interested in.

Remark 6. The three other nonlinear shrinkage formulas that do not fit into the mold of (18) are just elementary combinations of $\widehat{\varphi}_n^{\gamma}(\cdot)$ for two different γ functions.

Remark 7. It should also be pointed out that, apart from the two special cases $\gamma(x)=x^{\pm1}$, these two infinite families of loss functions can only be handled by using the new technique of angle estimation introduced in Section 3.3.

Table 5
Average losses computed for various estimators for $p = 100$ and $n = 200$. The first column gives the name of the loss function and the remaining columns give the average losses (over 1000 simulations) for a given estimator. In each row, the best (that is, smallest) number is in **bold face**.

Loss function	FSOPT	Identity	Sample	Linear	$\widehat{S}_n^{\circ}$	$\widehat{S}_n^{*}$	$\widehat{S}_n^{\circledast}$	$\widehat{S}_n^{\mathrm{LE}}$	$\widehat{S}_n^{\mathrm{FRÉ}}$	$\widehat{S}_n^{\mathrm{Q}}$	$\widehat{S}_n^{\mathrm{QINV}}$
Frobenius	5.76	14.64	14.75	7.39	**5.93**	7.75	6.45	6.33	6.04	16.08	8.23
Inverse Stein	0.15	0.33	0.71	0.18	**0.16**	0.22	0.17	0.17	0.16	0.46	0.22
Minimum Variance	1.10	2.72	2.75	1.37	**1.14**	1.16	1.15	1.16	1.15	1.17	1.36
Stein	0.15	0.69	0.31	0.29	0.21	**0.16**	0.17	0.17	0.19	0.22	0.51
Inverse Frobenius	0.05	0.14	0.84	0.10	0.07	**0.05**	0.06	0.05	0.07	0.07	0.13
Symmetrized Stein	0.33	1.02	1.02	0.47	0.37	0.37	**0.34**	0.34	0.35	0.69	0.74
Weighted Frobenius	0.23	1.02	0.50	0.38	0.32	**0.23**	0.25	0.26	0.28	0.34	0.74
Inverse Weighted Frob.	0.29	0.50	5.22	0.34	**0.30**	0.44	0.33	0.34	0.31	0.91	0.41
Log-Euclidian	0.29	0.86	0.76	0.43	0.33	0.32	0.30	**0.30**	0.31	0.60	0.64
Fréchet	0.29	0.77	0.59	0.37	0.30	0.35	0.30	0.30	**0.30**	0.70	0.50
Quadratic	0.29	4.21	1.01	1.29	1.00	0.46	0.65	0.67	0.80	**0.30**	2.93
Inverse Quadratic	0.26	0.50	9.45	0.38	0.50	1.10	0.68	0.74	0.58	2.63	**0.27**

6. Monte Carlo Simulations

The goal of this section is to illustrate on simulated data that there is generally great benefit in using the shrinkage estimator that is tailored to the loss function one has selected.

6.1. General setup

The population eigenvalues are distributed as follows: 20% are equal to 1, 40% are equal to 3, and 40% are equal to 10. This is a challenging problem originally introduced by [2]. We use the 12 loss functions from Tables 1–3. For each one, we compute the FSOPT estimator specific to the particular loss function, as well as all 7 bona fide shrinkage estimators presented in the paper. We use the same notation as [22]: $\widehat{S}_n^{\circ}$ is the estimator optimal with respect to Frobenius, Inverse Stein and Minimum Variance losses; $\widehat{S}_n^{*}$ is the one optimal with respect to Stein and Inverse Frobenius losses; and $\widehat{S}_n^{\circledast}$ the one optimal with respect to the Symmetrized Stein's loss. In addition, the identity matrix (rescaled to have same trace as the sample covariance matrix), the sample covariance matrix, and the linear shrinkage estimator of [18] are also computed for reference purposes. The results are averaged over 1000 simulations.

6.2. Nonsingular case

To produce the results of Table 5, the matrix dimension is $p = 100$ and the sample size is $n = 200$. In each row, the performance of the best bona fide estimator is printed in bold. One can see that the winner is always the estimator tailor-made for the loss function of the given row. Sometimes the difference with the other estimators is quite stark. Obviously, the FSOPT always dominates, but usually the excess loss of the best bona fide estimator is quite small. This finding reinforces the message that the asymptotically optimal estimators listed in the present paper perform as well as they ought to, even in finite samples.

Regarding the other (reference) estimators, linear shrinkage does better than the two ingredients that it interpolates, the scaled identity matrix and the sample covariance matrix, with respect to all but one of the 12 loss functions. This is good news because in theory its shrinkage intensity is optimized with respect to the Frobenius loss only. Linear shrinkage performs honorably across the board for such a simple estimator: it even manages to beat some nonlinear shrinkage estimators in almost every row, typically a couple of them. Needless to say, linear shrinkage never beats the nonlinear shrinkage formula optimized to the loss function in the given row, which shows that it 'leaves some money on the table' and that shrinking nonlinearly (in the appropriate way) delivers yet another round of improvement over and above linear shrinkage.

6.3. Comparison of shrinkage formulas

Confirming the ordering of Proposition 1, Fig. 3 gives further insight into the loss functions by showing how the 7 estimators shrink the sample eigenvalues in this case.

The Quadratic and the Inverse Quadratic shrinkage formulas stand out as 'outliers', as shown by Proposition 1. In Table 5, the estimators $\widehat{S}_n^{\mathrm{Q}}$ and $\widehat{S}_n^{\mathrm{QINV}}$ display erratic performances when measured against other loss functions than their own. The other estimators are better able to deliver respectable performance across foreign loss functions. The estimators $\widehat{S}_n^{\circ}$ and $\widehat{S}_n^{*}$ have strong backing, from the Minimum-Variance and Stein's loss, respectively; the Log-Euclidian estimator $\widehat{S}_n^{\mathrm{LE}}$ represents a'neutral' compromise that has strong foundations in the differential geometry of the manifold of tensors (a.k.a. positive definite matrices).

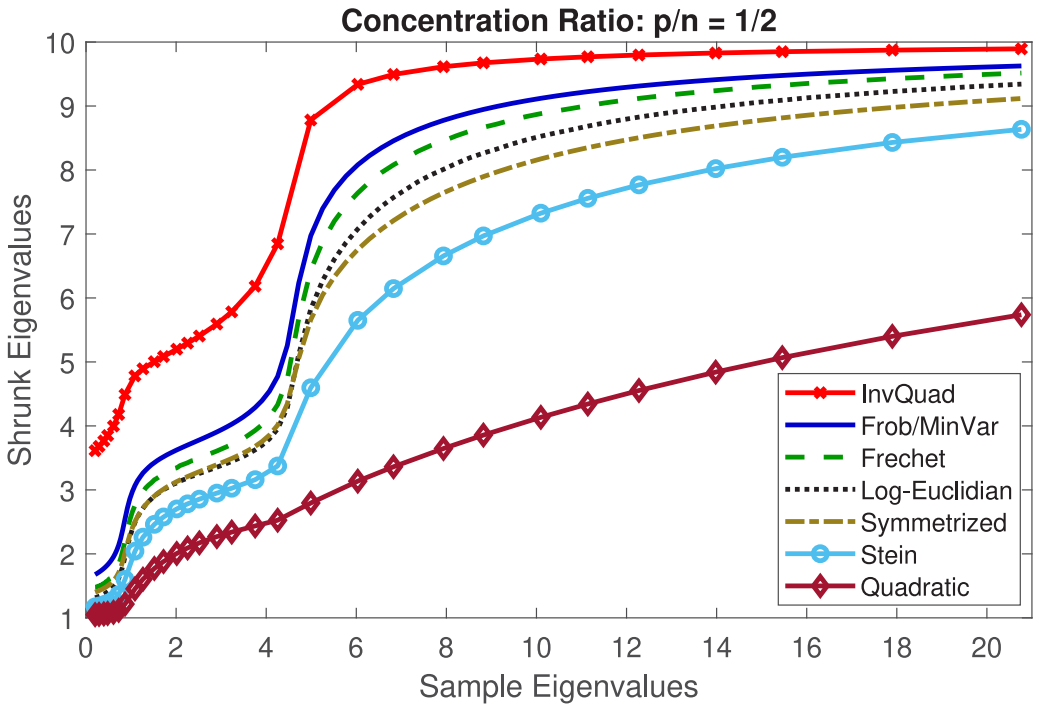

Fig. 3. Comparison of 7 nonlinear shrinkage formulas. For a sample eigenvalue on the x-axis, the value on the y-axis corresponds to the output of a given shrinkage formula.

Table 6
Average losses computed for various estimators for $p = 200$ and $n = 100$. The first column gives the name of the loss function and the remaining columns give the average losses (over 1000 simulations) for a given estimator. In each row, the best (that is, smallest) number is in **bold face**.

Loss function	FSOPT	Identity	Linear	$\widehat{S}_n^{\circ}$	$\widehat{S}_n^{*}$	$\widehat{S}_n^{\circledast}$	$\widehat{S}_n^{\mathrm{LE}}$	$\widehat{S}_n^{\mathrm{FRÉ}}$	$\widehat{S}_n^{\mathrm{Q}}$	$\widehat{S}_n^{\mathrm{QINV}}$
Frobenius	11.25	14.64	11.77	**11.36**	15.34	12.59	12.60	11.69	22.04	17.56
Inverse Stein	0.27	0.33	0.28	**0.28**	0.42	0.31	0.32	0.29	0.73	0.36
Minimum Variance	2.22	2.72	2.27	**2.23**	2.26	2.24	2.25	2.24	2.30	2.36
Stein	0.29	0.69	0.51	0.50	**0.30**	0.36	0.35	0.41	0.40	1.07
Inverse Frobenius	0.09	0.14	0.13	0.13	**0.09**	0.11	0.10	0.11	0.10	0.16
Symmetrized Stein	0.66	1.02	0.79	0.77	0.72	0.67	**0.66**	0.69	1.07	1.43
Weighted Frobenius	0.40	1.02	0.71	0.70	**0.41**	0.48	0.47	0.56	0.46	1.93
Disutility	0.45	0.50	0.46	**0.46**	0.69	0.50	0.52	0.47	1.29	0.53
Log-Euclidian	0.59	0.86	0.69	0.67	0.64	0.60	**0.59**	0.61	0.91	1.12
Fréchet	0.57	0.77	0.61	0.60	0.70	0.59	0.59	**0.58**	1.03	1.01
Quadratic	0.40	4.21	2.72	2.65	0.80	1.42	1.37	1.90	**0.49**	8.42
Inverse Quadratic	0.32	0.50	0.52	0.53	1.85	0.95	1.03	0.73	4.15	**0.32**

6.4. Singular case

Table 6 presents further results when $p = 200$ and $n = 100$.

Once again, the pattern is confirmed overall, except for one violation: $\widehat{S}_n^{\mathrm{LE}}$ beats $\widehat{S}_n^{\circledast}$ both 'home' and 'away': with respect to the Log-Euclidian loss and, unexpectedly, with respect to the Symmetrized Stein's loss also. (In other simulations not reported here, we double-checked that $\widehat{S}_n^{\circledast}$ does beat $\widehat{S}_n^{\mathrm{LE}}$ with respect to the Symmetrized Stein's loss when dimension is high enough, as implied by large-dimensional asymptotic theory.) Both of these estimators plow the same narrow but interesting field of estimators that are equivariant with respect to matrix inversion, so it is not completely surprising that the estimator that beats $\widehat{S}_n^{\circledast}$ on its home turf shares the same desirable property.

Remarks regarding the two simple estimators (scaled identity and linear shrinkage) essentially go in the same direction as in Section 6.2. We excluded the sample covariance matrix because it is not invertible, so most of the loss functions return $+\infty$.

6.5. Comparison with simpler alternatives: the matrix as a whole

We examine two alternative approaches that make compromises in order to obtain formulas that are simpler than the ones developed in this paper. On the one hand, linear shrinkage [18] compromises by forcing all eigenvalues to be shrunk

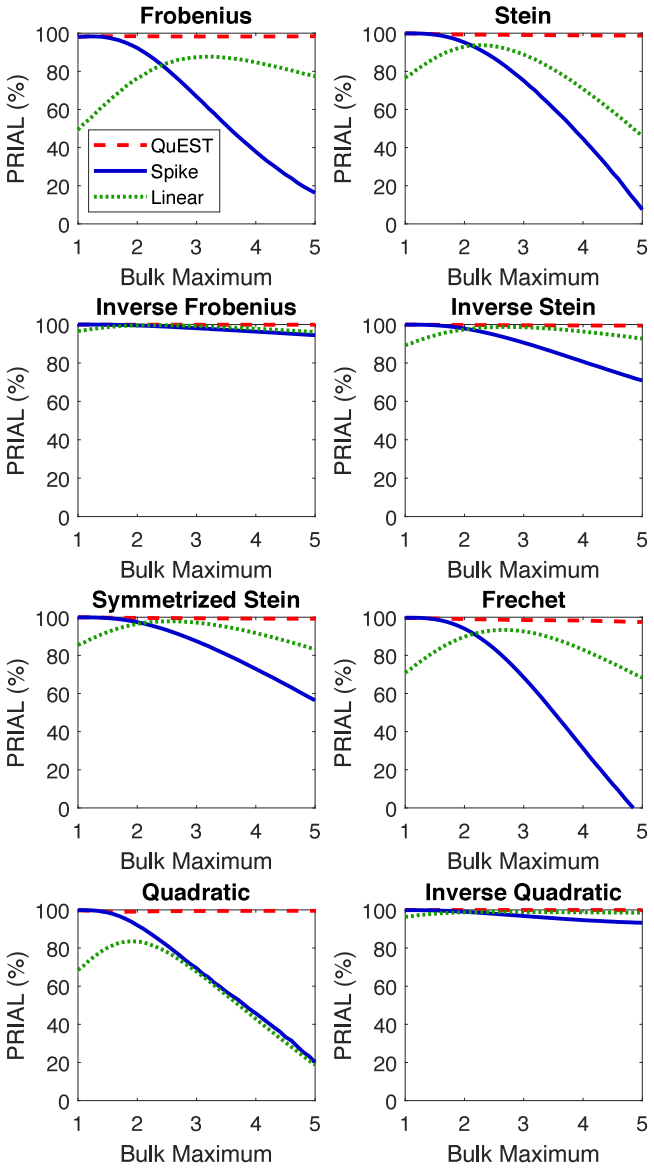

Fig. 4. PRIAL as a function of the spread of the bulk of the population eigenvalues for 3 bona fide estimators under 8 loss functions.

towards the same target with the same shrinkage intensity, and by considering only the Frobenius loss; on the other hand, the spiked-model approach [5] compromises by assuming that the bulk of the population eigenvalues (meaning: all of them except for a vanishing fraction of them, called spikes) are equal.

In this part, we take both linear shrinkage and spike shrinkage 'outside of their comfort zone' by considering 8 different loss functions, on the one hand, and by considering specifications where the bulk of the population eigenvalues can be distinct, on the other hand. Most applied researchers will be interested to know how robust the simplified formulas are against violations of the framework under which they have been derived.

The 8 loss functions that we consider are the ones in the intersection of the 12 considered in this paper with the 18 for which [5] provide closed-form spike shrinkage: (1) Frobenius, (2) Stein, (3) Inverse Frobenius, 4) Inverse Stein, (5) Symmetrized Stein, (6) Fréchet, (7) Quadratic, and (8) Inverse Quadratic. The code for spike shrinkage was taken directly from [4]. As far as the population eigenvalues are concerned, the initial specification is to have a single spike at 10, and the $p-1$ bulk eigenvalues equal to 1. From this base, we will allow for heterogeneity in the bulk by keeping half of the bulk equal to 1, while setting the other half equal to $\overline{\tau} \in [1,5]$. It is only fair to allow bulk eigenvalues to be distinct: after all, this is the generic case in real-world applications. We put 100 eigenvalues in the bulk, plus (as mentioned above) a single spike, for a total of $p = 101$ eigenvalues. We take the (limiting) concentration ratio to be $c = 1/2$, which implies $n = 202$.

Fig. 4 displays the Percentage Relative Improvement in Average Loss (PRIAL):

$$\text{PRIAL}\left(\mathcal{L}_n^i, \widetilde{S}_n\right) := \frac{\mathbb{E}\left[\mathcal{L}_n^i\left(S_n, \Sigma_n\right)\right] - \mathbb{E}\left[\mathcal{L}_n^i\left(\widetilde{S}_n, \Sigma_n\right)\right]}{\mathbb{E}\left[\mathcal{L}_n^i\left(S_n, \Sigma_n\right)\right] - \mathbb{E}\left[\mathcal{L}_n^i\left(\widehat{S}_n^{*,i}, \Sigma_n\right)\right]} \times 100\% \,, \tag{32}$$

where $\mathcal{L}_n^i$ denotes one of the eight loss functions listed above, $\widehat{S}_n^{*,i}$ denotes the FSOPT estimator tailored to each specific loss function as per Tables 1–3, $\widetilde{S}_n$ denotes the estimator under consideration (whether linear shrinkage, spike shrinkage, or nonlinear shrinkage), and the expectation is approximated by the average of 1000 Monte Carlo simulations. By construction, the PRIAL of the sample covariance matrix is 0% whereas the PRIAL of the FSOPT estimator is 100%. Therefore, the PRIAL measures how much of the potential for improvement relative to the sample covariance matrix is attained by a given estimator $\widetilde{S}_n$.

One can see that, even though the dimension is not overly large ($p \approx 100$), nonlinear shrinkage captures nearly 100% of the potential improvement with respect to all loss functions, regardless of the spread of the bulk eigenvalues. Linear shrinkage has more of a mixed performance, but still generally manages a PRIAL of at least 50%. It beats the sample covariance matrix for all 8 loss functions, which shows that its attractiveness extends far beyond the Frobenius loss. It also always (weakly) beats spike shrinkage as long as $\overline{\tau} \geq 2.5$. Note that linear shrinkage is the only estimator that keeps the same formula throughout, so it is 'fighting with one hand tied behind the back' when it has to compete under the 7 loss functions different from Frobenius loss.

As expected, the performance of spike shrinkage is near-perfect when its specification matches reality ($\overline{\tau} = 1$: all bulk eigenvalues are equal), but it monotonically degrades as soon as the bulk population eigenvalues become heterogeneous. This drop in performance is not so pronounced with the Inverse Frobenius, Inverse Stein and Inverse Quadratic losses, but it is very pronounced with the 5 other loss functions. There is even a case ($\overline{\tau} = 5$ and Fréchet loss) where spike shrinkage underperforms the sample covariance matrix, which results in a negative PRIAL. This is a result that should be expected purely from theory: Unlike linear and nonlinear shrinkage, spike shrinkage can actually be worse than the sample covariance matrix even in the large-dimensional asymptotic limit.

Compared to the two simpler alternatives, nonlinear shrinkage does better across the board. In particular, there are scenarios where the optimal nonlinear shrinkage formula is (nearly) linear; and, even then, nonlinear shrinkage performs just as well as linear shrinkage, for all practical purposes. Similarly, there are scenarios where the spiked covariance model holds perfectly true (bulk maximum equal to one); and, even then, nonlinear shrinkage performs just as well as spike shrinkage, for all practical purposes.

The overall conclusion is that, among the simpler formulas, linear shrinkage can 'leave some money on the table' when the optimal shrinkage is highly nonlinear whereas spike shrinkage is vulnerable to the risk that its stringent specification of bulk-eigenvalue equality is violated by reality. Only the full-blown nonlinear shrinkage formulas derived in this paper avoid both pitfalls and deliver state-of-the-art enhancement of the sample covariance matrix across the board.

6.6. Comparison with simpler alternatives: focus on the spike

There might be a perception that nonlinear shrinkage does better on the bulk, whereas spike shrinkage does better on the spike. This is hard to justify formally, as the loss functions used in the spike literature pertain to the whole covariance matrix, placing no special ()over-)weight on the spike. Nonetheless, in Table 7, which appears in the supplementary material, we isolate the spike's contribution to the overall loss, for the 8 loss functions from Section 6.5. This is easy to do because every one of these loss functions can be decomposed as a sum of individual contributions attributable to the p eigenvalues. Fig. 5 displays the corresponding Monte Carlo simulation results (obviously in terms of risk instead of loss now).

The performance of the sample covariance matrix can be so erratic as to not even be on the same scale as the other estimators in many scenarios. The risk contribution of the FSOPT is zero by construction. The pattern is that spike shrinkage and nonlinear shrinkage estimate the spike equally well when the spike model's assumption of equal bulk eigenvalues is satisfied, or nearly satisfied. However, as the bulk spreads out, nonlinear shrinkage estimates the spike more accurately than spike shrinkage for all 8 loss functions where the two methods overlap.

7. Conclusion

In this paper, we have

- developed a new estimator of the angle between any sample eigenvector and any population eigenvector by exploiting a sophisticated equation from random matrix theory (RMT);
- doubled the number of loss functions that can be handled from 6 to 12 (compared to related earlier work), which can only be achieved by the new technique of angle estimation;
- proposed a classification of loss functions by their finite-sample optimal shrinkage formulas;
- increased the number of asymptotically optimal nonlinear shrinkage formulas from 3 to 7 (compared to related earlier work);
- established an ordering of the nonlinear shrinkage formulas (from largest to smallest);

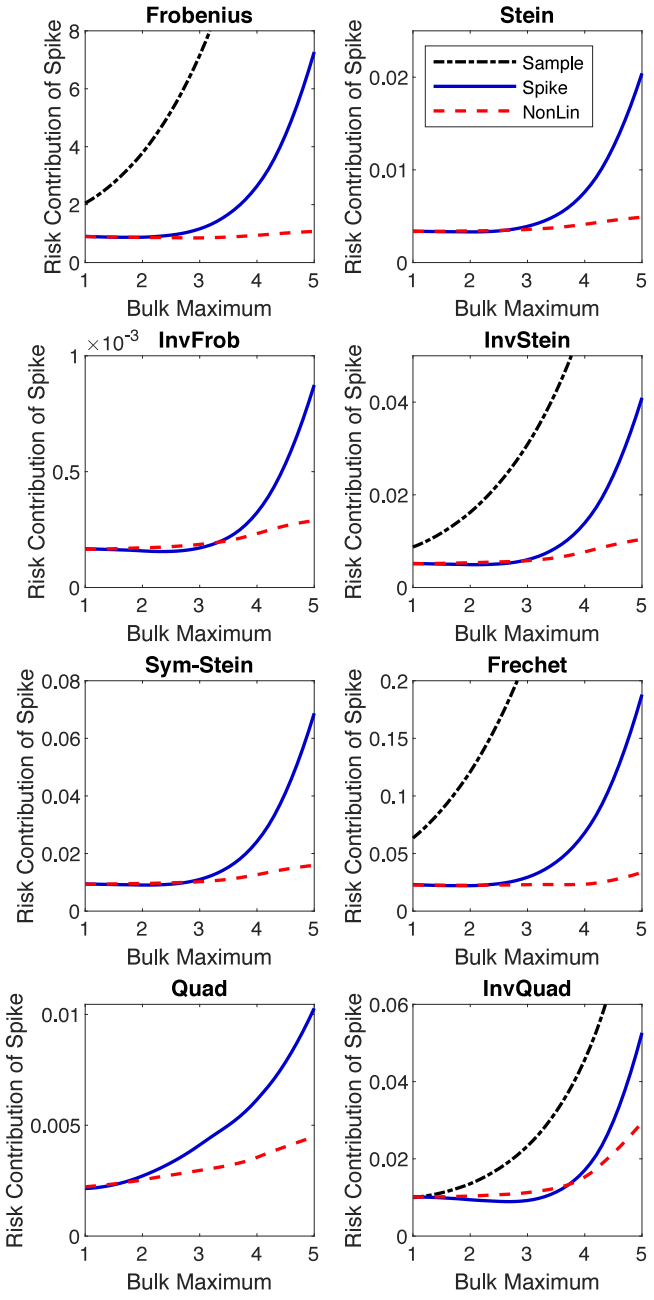

Fig. 5. Normalized contribution of the spike to the overall risk of the covariance matrix for 8 loss functions, as a function of the spread of the bulk of population eigenvalues. The setup is the same as in Fig. 4, except that we run 10,000 simulations instead of 1000 simulations.

- delivered two infinite families of loss functions and their (correspondingly infinite family of) optimal nonlinear shrinkage formulas, which can only be achieved by the new technique of angle estimation;
- and introduced a new loss function founded on the economic concept of utility maximization.

As a simpler alternative approach, Donoho, Gavish, and Johnstone [5] consider a ménagerie of 26 loss functions under the spiked covariance model of [13]. The key distinction in this model is between the bulk, which is comprised by eigenvalues packed shoulder-to-shoulder like sardines, and the spikes, which are a few select eigenvalues large enough to separate from the bulk. Donoho et al. [5] treat the spikes carefully, but they just collapse the bulk. This approach is perfectly legitimate under the assumption that they make, namely, that all bulk population eigenvalues are equal. However, in many applications, such an assumption is unrealistic or may not be known to hold. In the general case,

not all bulk population eigenvalues are equal, so valuable information can be gleaned from the angle between sample and population eigenvectors, and from applying differentiated shrinkage inside the bulk. Monte Carlo simulations show that the resulting nonlinear shrinkage performs, for all practical purposes, just as well as spike shrinkage when all bulk population eigenvalues are equal, but performs often much better when they are not. Therefore, in this context, the KISS (Keep it simple, statistician!) principle does not seem to benefit applied researchers: By upgrading from spike shrinkage to full-blown nonlinear shrinkage, they have, for all practical purposes, nothing to lose but much to gain. In addition, at least currently, spike shrinkage, unlike nonlinear shrinkage, is only available for the case where the dimension is smaller than the sample size, which limits practical applications.

Having said this, Donoho et al. [5] roll out a clever technology that convincingly documents three closely interrelated facts that have not garnered sufficient attention in this field:

1. The choice of loss function has a profound effect on optimal estimation.
2. Eigenvalue inconsistency: The sample eigenvalues are spread, biased, and shifted away from their theoretical (population) counterparts by an asymptotically predictable amount.
3. Eigenvector inconsistency: The angles between the sample eigenvectors and the corresponding population eigenvectors have nonzero asymptotic limits.

Such fundamental truths need to be hammered in again and again, in every possible way.

Finally, we may say a word about the choice of loss function. 12 of them have been solved already, yielding seven different nonlinear shrinkage formulas, in addition to the two infinite families, which should be more than enough to satisfy any reasonable need. By definition, it is the duty of the end-user to pick the loss function, but perhaps some light-touch guidance can help orient readers through a forest with so many trees. For anyone interested in using a covariance matrix estimator to minimize variance, risk, or noise in any sense, certainly the Minimum Variance loss function is the appropriate one; an additional advantage is that for this loss function a new technology has arisen that is no more complex than kernel density estimation, and so is extremely fast and scalable to ultra-high dimensions [23]. For researchers concerned with the decision-theoretic aspects of the problem, a loss function based on the Kullback–Leibler divergence (also called relative entropy), such as Stein's loss, is the natural candidate. For other applications, such as fMRI tensors, where it is important to regard eigenvalues close to zero as being 'as distant' as eigenvalues close to infinity, then the Log-Euclidian loss function is well suited: It appears a good compromise because it produces shrunken eigenvalues that lie in between the ones from the Minimum-Variance loss and the ones from Stein's loss. For all other categories of problems, integrating over all possible permutations/unknown directions, and using some approximations from random matrix theory, as per Section 4 of [8] may be used to zoom in on a specific loss function attuned to the situation at hand.

Acknowledgments

We thank the Editor-in-Chief, Dietrich von Rosen, and an anonymous referee for helpful comments that have enhanced the exposition of the paper.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2021.104796.

References

[1] V. Arsigny, P. Fillard, X. Pennec, N. Ayache, Log-Euclidean metrics for fast and simple calculus on diffusion tensors, Magn. Reson. Med. 56 (2) (2006) 411–421.
[2] Z.D. Bai, J.W. Silverstein, No eigenvalues outside the support of the limiting spectral distribution of large-dimensional random matrices, Ann. Probab. 26 (1) (1998) 316–345.
[3] P.J. Bickel, E. Levina, Regularized estimation of large covariance matrices, Ann. Statist. 36 (1) (2008) 199–227.
[4] D.L. Donoho, M. Gavish, I.M. Johnstone, Code supplement to "Optimal shrinkage of eigenvalues in the spiked covariance model", 2016, Available online at http://purl.stanford.edu/xy031gt1574.
[5] D.L. Donoho, M. Gavish, I.M. Johnstone, Optimal shrinkage of eigenvalues in the spiked covariance model, Ann. Statist. 46 (4) (2018) 1742–1778.
[6] D. Dowson, B. Landau, The Fréchet distance between multivariate normal distributions, J. Multivariate Anal. 12 (3) (1982) 450–455.
[7] N. El Karoui, Spectrum estimation for large dimensional covariance matrices using random matrix theory, Ann. Statist. 36 (6) (2008) 2757–2790.
[8] R.F. Engle, O. Ledoit, M. Wolf, Large dynamic covariance matrices, J. Bus. Econom. Statist. 37 (2) (2019) 363–375.
[9] M. Ghosh, B. Sinha, Inadmissibility of the best equivariant estimators of the variance-covariance matrix, the precision matrix and the generalized variance under entropy loss, Statist. Decisions 5 (3–4) (1987) 201–228.
[10] L. Haff, Estimation of the inverse covariance matrix: Random mixtures of the inverse Wishart matrix and the identity, Ann. Statist. (1979) 1264–1276.
[11] L. Haff, An identity for the Wishart distribution with applications, J. Multivariate Anal. 9 (4) (1979) 531–544.
[12] W. James, C. Stein, Estimation with quadratic loss, in: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, 1961, pp. 361–380.
[13] I.M. Johnstone, On the distribution of the largest eigenvalue in principal component analysis, Ann. Statist. 29 (2) (2001) 295–327.
[14] W. Jun, On High-Dimensional Covariance Matrices Estimation (Ph.D. thesis), National University of Singapore, Department of Statistics and Applied Probability, 2017, Available online at http://scholarbank.nus.edu.sg/handle/10635/135450.

[15] T. Kubokawa, Y. Konno, Estimating the covariance matrix and the generalized variance under a symmetric loss, Ann. Inst. Statist. Math. 42 (2) (1990) 331–343.
[16] S. Kullback, R.A. Leibler, On information and sufficiency, Ann. Math. Stat. 22 (1) (1951) 79–86.
[17] O. Ledoit, S. Péché, Eigenvectors of some large sample covariance matrix ensembles, Probab. Theory Related Fields 150 (1–2) (2011) 233–264.
[18] O. Ledoit, M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, J. Multivariate Anal. 88 (2) (2004) 365–411.
[19] O. Ledoit, M. Wolf, Nonlinear shrinkage estimation of large-dimensional covariance matrices, Ann. Statist. 40 (2) (2012) 1024–1060.
[20] O. Ledoit, M. Wolf, Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions, J. Multivariate Anal. 139 (2) (2015) 360–384.
[21] O. Ledoit, M. Wolf, Numerical implementation of the QuEST function, Comput. Statist. Data Anal. 115 (2017) 199–223.
[22] O. Ledoit, M. Wolf, Optimal estimation of a large-dimensional covariance matrix under Stein's loss, Bernoulli 24 (4B) (2018) 3791–3832.
[23] O. Ledoit, M. Wolf, Analytical nonlinear shrinkage of large-dimensional covariance matrices, Ann. Statist. 40 (5) (2020) 3043–3065.
[24] P.L. Leung, R.J. Muirhead, Estimation of parameter matrices and eigenvalues in MANOVA and canonical correlation analysis, Ann. Statist. 15 (4) (1987) 1651–1666.
[25] H. Markowitz, Portfolio selection, J. Finance 7 (1952) 77–91.
[26] J.B. Selliah, Estimation and Testing Problems in a Wishart Distribution (Ph.D. thesis), Stanford University, Department of Statistics, 1964.
[27] D. Sharma, K. Krishnamoorthy, Empirical Bayes estimators of normal covariance matrix, SankhyÃ: Indian J. Statist., Ser. A (1985) 247–254.
[28] J.W. Silverstein, Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices, J. Multivariate Anal. 55 (1995) 331–339.
[29] J.W. Silverstein, S.I. Choi, Analysis of the limiting spectral distribution of large-dimensional random matrices, J. Multivariate Anal. 54 (1995) 295–309.
[30] C. Stein, Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, in: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 1956, pp. 197–206.
[31] C.M. Stein, Multivariate analysis I, Technical Report No. 42, Department of Statistics, Stanford University, 1969, (Notes prepared by Morris L. Eaton.).
[32] C. Stein, Estimation of a covariance matrix, in: Rietz Lecture, 39th Annual Meeting IMS, Atlanta, Georgia, 1975.
[33] C. Stein, Lectures on the theory of estimation of many parameters, J. Math. Sci. 34 (1) (1986) 1373–1403.
[34] E.P. Wigner, Characteristic vectors of bordered matrices with infinite dimensions, Ann. of Math. 62 (3) (1955) 548–564.

Further reading

[1] O. Ledoit, M. Wolf, The power of (non-)linear shrinking: A review and guide to covariance matrix estimation, J. Financ. Econom. (2020) Available at http://dx.doi.org/10.1093/jjfinec/nbaa007.
[2] G. Pan, Comparison between two types of large sample covariance matrices, Ann. Inst. H. Poincaré Probab. Statist. 50 (2) (2014) 655–677.
[3] J.W. Silverstein, Z.D. Bai, On the empirical distribution of eigenvalues of a class of large-dimensional random matrices, J. Multivariate Anal. 54 (1995) 175–192.