Ledoit, Olivier; Wolf, Michael

Working Paper

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

Working Paper, No. 327

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

Suggested Citation: Ledoit, Olivier; Wolf, Michael (2019) : Shrinkage estimation of large covariance matrices: Keep it simple, statistician?, Working Paper, No. 327, University of Zurich, Department of Economics, Zurich, http://dx.doi.org/10.5167/uzh-172202

This Version is available at:
http://hdl.handle.net/10419/201532

University of Zurich

Department of Economics

Working Paper Series

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

Working Paper No. 327

Shrinkage Estimation of Large Covariance Matrices: Keep it Simple, Statistician?

Olivier Ledoit and Michael Wolf

July 2019

Shrinkage Estimation of Large Covariance Matrices: Keep it Simple, Statistician?

Olivier Ledoit*
Department of Economics
University of Zurich
CH-8032 Zurich, Switzerland
olivier.ledoit@econ.uzh.ch

Michael Wolf
Department of Economics
University of Zurich
CH-8032 Zurich, Switzerland
michael.wolf@econ.uzh.ch

July 2019

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 introduce a broad family of covariance matrix estimators that can handle all regular functional transformations of the population covariance matrix under large-dimensional asymptotics. We solve the problem of optimal covariance matrix estimation under a variety of loss functions motivated by statistical precedent, probability theory, and differential geometry. The key statistical 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 compare our methodology to two simpler ones from the literature, linear shrinkage and shrinkage based on the spiked covariance model, via both Monte Carlo simulations and an empirical application.

KEY WORDS: Large-dimensional asymptotics, random matrix theory, rotation equivariance.

JEL CLASSIFICATION NOS: C13.

*Second affiliation: AlphaCrest Capital, 120 W 45th St, New York NY 10036, USA.

1 Introduction

Ever since Stein (1956) proved that the usual estimator of the mean 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 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 by us so far within the framework of large-dimensional asymptotics, yielding a grand total of three different optimal nonlinear shrinkage formulas. However, the third one is just a geometric average of the first two, so in essence we have only two 'views' into the entrails of the population covariance matrix. By this we mean how sample eigenvectors recombine with (a function of) the population covariance matrix.

This paper delivers the technology to get an infinite number of views into the population covariance matrix, in the absence of strict assumptions. To demonstrate the power of our method, we select four functional views that are potentially attractive to applied researchers and have not been explorable before. They tie up nicely with a number of pre-existing loss functions that have been promoted by statisticians for decision-theoretical estimation of the covariance matrix, as well as metrics defined on the space of symmetric positive-definite matrices by mathematicians. In order to achieve this degree of generality, we identify a formula from random matrix theory (RMT) that enables us to estimate the angle of any sample eigenvector with any population eigenvector, in the large-dimensional asymptotic limit. It will be useful to give a brief review of the relevant literature before starting to develop our methodology.

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. The 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 Stein (1975, 1986): (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 (1955), 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 (Stein, 1969, 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 (Bickel and Levina, 2008) 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 (Ledoit and Wolf, 2004) violates point (iii) because it can only modify the eigenvalues of the sample covariance matrix through a linear transformation. The spiked covariance model of Johnstone (2001) violates point (i) because it assumes that all population eigenvalues are equal to each other, except for a vanishingly small proportion of them.

By contrast, the present paper inscribes itself in a strand of literature called *nonlinear shrinkage* (Ledoit and Wolf, 2012, 2015, 2018b) that does not compromise on any of these three points, and so remains in line with Stein's original ambitious paradigm. A basic ingredient is consistent estimation of the eigenvalues of the population covariance matrix. This was not even deemed possible until El Karoui (2008) proved otherwise. Since then, it has been more of a discussion of which estimation scheme to use, such as El Karoui's own numerical procedure or a more modern approach based on supersymmetry (Jun, 2017); in this paper, we use the QuEST function of Ledoit and Wolf (2015).

In recent related work, the spiked covariance model of Johnstone (2001) has been used by Donoho et al. (2018) to derive shrinkage covariance matrix estimators for a *ménagerie* of 26 different loss functions. They 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 Donoho et al. (2018) 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 presents an empirical application to real data. Section 8 concludes. An appendix collects various mathematical results to keep the presentation in the main paper compact.

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 sample covariance matrix is $S := Y'Y/n$. Its spectral decomposition is $S =: U\Lambda U'$, where Λ is a diagonal matrix, and U is orthogonal. Let $\Lambda =: \mathsf{Diag}(\boldsymbol{\lambda})$ where $\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_p)'$, 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 $S = \sum_{i=1}^{p} \lambda_i \cdot u_i u_i'$. Note that it holds similarly $\Sigma = \sum_{i=1}^{p} \tau_i \cdot v_i v_i'$.

Definition 2.1. *We consider rotation-equivariant covariance matrix estimators of the type* $\widetilde{S} := U\widetilde{D}U'$*, where* $\widetilde{D}$ *is a diagonal matrix:* $\widetilde{D} := \mathsf{Diag}(\widetilde{d}_1, \ldots, \widetilde{d}_p)$*.*

This class assumes no *a priori* information about the orientation of the orthonormal basis of (unobservable) population covariance matrix eigenvectors; this is different from the sparsity literature, which requires *a priori* knowledge of an orthonormal basis in which most covariances are zero. For many loss functions, there exists a finite-sample optimal (FSOPT) estimator in this class, of the form

$$\widetilde{S} := \sum_{i=1}^{p} \widetilde{d}_i \cdot u_i u_i', \qquad \text{with} \qquad \widetilde{d}_i := \gamma^{-1}\left[u_i' \gamma(\Sigma) u_i\right] \qquad \forall i = 1, \ldots, p\ , \tag{2.1}$$

where γ denotes some smooth invertible function mapping of $(0, +\infty)$ onto $\mathbb{R}$. As is standard, applying a univariate function γ to a diagonalizable matrix means preserving its eigenvectors and applying γ to each eigenvalue individually; for example, $\log(\Sigma) := \sum_{i=1}^{p} \log(\tau_i) \cdot v_i v_i'$. When we speak of *functional shrinkage*, or of a *view into the entrails* of the population covariance matrix, we mean which function γ to use.

2.2 A Brief Summary of Known Results on Nonlinear Shrinkage

In our previous work so far, only six loss functions have been solved in the rotation-equivariant framework of Assumption 1 and Definition 2.1. In the second column of Table 2.1, 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 $\lambda_i = u_i' S u_i$ $(i = 1, \ldots, p)$ 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: $\|M\|_{\mathrm{F}} := \sqrt{\mathsf{Tr}[MM']}$.

Table 2.1 shows that the six loss functions really only yield three different nonlinear shrinkage formulas. The first two are of the type (2.1), 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 2.2 presents two of them.

Name	Stylized Loss Function	Reference	Shrinkage
Frobenius	$\left\|\Sigma - \widetilde{S}\right\|_{\mathrm{F}}$	Leung and Muirhead (1987)	$u_i'\Sigma u_i$
Inverse Stein	$\mathsf{Tr}\left[\Sigma\widetilde{S}^{-1}\right] - \log\left[\det\left(\Sigma\widetilde{S}^{-1}\right)\right]$	Ghosh and Sinha (1987)	$u_i'\Sigma u_i$
Minimum Variance	$\mathsf{Tr}\left[\widetilde{S}^{-1}\Sigma\widetilde{S}^{-1}\right] \Big/ \left(\mathsf{Tr}\left[\widetilde{S}^{-1}\right]\right)^2$	Engle et al. (2019)	$u_i'\Sigma u_i$
Stein	$\mathsf{Tr}\left[\Sigma^{-1}\widetilde{S}\right] - \log\left[\det\left(\Sigma^{-1}\widetilde{S}\right)\right]$	James and Stein (1961)	$\frac{1}{u_i'\Sigma^{-1}u_i}$
Inverse Frobenius	$\left\|\Sigma^{-1} - \widetilde{S}^{-1}\right\|_{\mathrm{F}}$	Haff (1979a)	$\frac{1}{u_i'\Sigma^{-1}u_i}$
Symmetrized Stein	$\mathsf{Tr}\left[\Sigma^{-1}\widetilde{S} + \Sigma\widetilde{S}^{-1}\right]$	Kubokawa and Konno (1990)	$\sqrt{\frac{u_i'\Sigma u_i}{u_i'\Sigma^{-1}u_i}}$

Table 2.1: Existing set of finite-sample optimal (FSOPT) nonlinear shrinkage formulas.

Name	Stylized Loss Function	Reference	Shrinkage
Weighted Frobenius	$\mathsf{Tr}\left[\left(\widetilde{S} - \Sigma\right)^2\Sigma^{-1}\right]$	Sharma and Krishnamoorthy (1985)	$\frac{1}{u_i'\Sigma^{-1}u_i}$
Disutility	$\mathsf{Tr}\left[\left(\widetilde{S}^{-1} - \Sigma^{-1}\right)^2\Sigma\right]$	Appendix A	$u_i'\Sigma u_i$

Table 2.2: Two more loss functions leading to existing nonlinear shrinkage formulas.

The second loss function is new. It is derived from the Sharma and Krishnamoorthy (1985) loss in the same way that the Inverse Frobenius loss is derived from the Frobenius loss, or that the Inverse Stein's loss of Ghosh and Sinha (1987) 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 Markowitz (1952) in large dimensions, as argued in Appendix A (hence the name *dis*utility). 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 three 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 $\gamma(x) = x^{\pm 1}$ 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.

Type of Loss	Loss Function	Reference	Shrinkage
Log-Euclidian	$\left\|\log(\Sigma) - \log(\widetilde{S})\right\|_{\mathrm{F}}$	Arsigny et al. (2006)	$\exp[u_i' \log(\Sigma) u_i]$
Fréchet	$\left\|\Sigma^{1/2} - \widetilde{S}^{1/2}\right\|_{\mathrm{F}}$	Dowson and Landau (1982)	$\left(u_i' \Sigma^{1/2} u_i\right)^2$
Quadratic	$\left\|\Sigma^{-1}\widetilde{S} - \mathbb{I}\right\|_{\mathrm{F}}$	$L^{F,3}$ in Donoho et al. (2018)	$\frac{u_i'\Sigma^{-1}u_i}{u_i'\Sigma^{-2}u_i}$
Inverse Quadratic	$\left\|\widetilde{S}^{-1}\Sigma - \mathbb{I}\right\|_{\mathrm{F}}$	$L^{F,4}$ in Donoho et al. (2018)	$\frac{u_i'\Sigma^{2}u_i}{u_i'\Sigma u_i}$

Table 2.3: New set of finite-sample optimal (FSOPT) nonlinear shrinkage formulas.

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 Selliah (1964, Section 2.2.4) and Haff (1979b, loss function L_2). 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 estimators in the fourth column can be found in Appendix B.

There exists a well-established ordering of the seven nonlinear shrinkage formulas.

Proposition 2.1. *Under Assumption 1, with probability one, for all* $i = 1, \ldots, p$,

$$\frac{u_i'\Sigma^2 u_i}{u_i'\Sigma u_i} > u_i'\Sigma u_i > \left(u_i'\sqrt{\Sigma}u_i\right)^2 > \exp\left[u_i'\log\left(\Sigma\right)u_i\right] > \frac{1}{u_i'\Sigma^{-1}u_i} > \frac{u_i'\Sigma^{-1}u_i}{u_i'\Sigma^{-2}u_i} \tag{2.2}$$

$$\left(u_i'\sqrt{\Sigma}u_i\right)^2 > \sqrt{\frac{u_i'\Sigma u_i}{u_i'\Sigma^{-1}u_i}} > \frac{1}{u_i'\Sigma^{-1}u_i} \tag{2.3}$$

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

2.5 Preview of General Result

FSOPT estimators of the form (2.1) 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 somewhere? 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: $u_i'\gamma(\Sigma)u_i$, for $i = 1, \ldots, p$, which is much more manageable. When the matrix dimension p is large, it is possible to approximate these quantities by the general formula:

$$\boxed{u_i'\gamma(\Sigma)u_i \approx \frac{1}{p}\sum_{j=1}^p \gamma(\widehat{\tau}_j)\cdot\left\{\frac{\frac{p}{n}\,\lambda_i\,\widehat{\tau}_j}{\left|\widehat{\tau}_j\left[1-\frac{p}{n}-\frac{p}{n}\,\lambda_i\,\breve{m}_{n,p}^{\widehat{\boldsymbol{\tau}}}(\lambda_i)\right]-\lambda_i\right|^2}\right\}}, \tag{2.4}$$

where $\widehat{\boldsymbol{\tau}} \coloneqq (\widehat{\tau}_1, \ldots, \widehat{\tau}_j)'$ is an estimator of the population eigenvalues, and $\breve{m}_{n,p}^{\widehat{\boldsymbol{\tau}}}(x)$ is the complex-valued function of real argument due to Ledoit and Wolf (2015, Section 2). Formula (2.4) generates *bona fide* covariance matrix estimators of the type (2.1) for all the loss functions in Table 2.3 by setting $\gamma(x)$ equal to $\log(x)$, $\sqrt{x}$, x^{-2}, or x^2. Given that $u_i'\gamma(\Sigma)u_i = \frac{1}{p}\sum_{j=1}^p \gamma(\tau_j)\cdot\left\{p(u_i'v_j)^2\right\}$, the term between curly brackets in (2.4) is simply an estimator of the dimension-normalized squared dot product of the ith sample eigenvector with the jth population eigenvector.

3 Large-Dimensional Asymptotic Framework

We now move on to formally establishing that plugging the approximation (2.4) into the generic nonlinear shrinkage formula (2.1) 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 some earlier paper such as, for example, Ledoit and Wolf (2018b, Section 3.1), 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.

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 $\boldsymbol{\tau}_n := (\tau_{n,1}, \ldots, \tau_{n,p})'$ 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. $\mathsf{Supp}(H)$, 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 $\{\tau_{n,1}, \ldots, \tau_{n,p}\}$ for large n.

Note that this includes Johnstone's (2001) spiked covariance model as a special case where the limiting population spectral distribution H is a step function with a single step.

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.*

The sample covariance matrix is defined as $S_n := n^{-1} Y_n' Y_n = n^{-1} \sqrt{\Sigma_n} X_n' X_n \sqrt{\Sigma_n}$. It admits a spectral decomposition $S_n =: U_n \Lambda_n U_n'$, where Λ_n is a diagonal matrix, and U_n is an orthogonal matrix: $U_n U_n' = U_n' U_n = \mathbb{I}_n$, where $\mathbb{I}_n$ (in slight abuse of notation) denotes the identity matrix of dimension $p \times p$. Let $\Lambda_n := \mathsf{Diag}(\boldsymbol{\lambda}_n)$ where $\boldsymbol{\lambda}_n := (\lambda_{n,1}, \ldots, \lambda_{n,p})'$. We can assume 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 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, due to Silverstein (1995): for all z in $\mathbb{C}^+$, the half-plane of complex numbers with strictly positive imaginary part,

$m \coloneqq m_{\mathrm{F}}(z)$ is the unique solution in the set $\left\{m \in \mathbb{C} : -\frac{1-c}{z} + cm \in \mathbb{C}^+\right\}$ to the equation

$$m = \int \frac{1}{\tau\big[1 - c - c\,z\,m\big] - z}\, dH(\tau) \ , \tag{3.1}$$

where m_{F} denotes the Stieltjes (1894) transform of F, whose standard definition is:

$$\forall z \in \mathbb{C}^+ \qquad m_{\mathrm{F}}(z) \coloneqq \int \frac{1}{\lambda - z}\, dF(\lambda) \ .$$

The Stieltjes transform admits a well-known inversion formula:

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

if G is continuous at both a and b. Although the Stieltjes transform of F, m_{F}, is a function whose domain is the upper half of the complex plane, it admits an extension to the real line, since Silverstein and Choi (1995) show that: $\forall x \in (0, +\infty)$, $\lim_{z \in \mathbb{C}^+ \to x} m_{\mathrm{F}}(z) =: \breve{m}_{\mathrm{F}}(x)$ exists and is continuous. The imaginary part of $\breve{m}_{\mathrm{F}}$ is the derivative of F, up to rescaling by π; therefore, (3.1) 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 $\mathsf{Supp}(F)$ is the union of a finite number $\kappa \geq 1$ of compact intervals: $\mathsf{Supp}(F) = \bigcup_{k=1}^{\kappa} [a_k, b_k]$, where $0 < a_1 < b_1 < \cdots < a_\kappa < b_\kappa < \infty$.

Definition 3.1 (Rotation-Equivariant Estimators)**.** *We consider covariance matrix estimators of the type $\widetilde{S}_n \coloneqq U_n \widetilde{D}_n U_n'$, where $\widetilde{D}_n$ is a diagonal matrix: $\widetilde{D}_n \coloneqq \mathsf{Diag}(\widetilde{\varphi}_n(\lambda_{n,1}) \ldots, \widetilde{\varphi}_n(\lambda_{n,p}))$, 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 $\mathsf{Supp}(F)$ and continuously differentiable on $\bigcup_{k=1}^{\kappa} [a_k, b_k]$ such that $\widetilde{\varphi}_n(x) \xrightarrow{\text{a.s.}} \widetilde{\varphi}(x)$ for all $x \in \mathsf{Supp}(F)$. Furthermore, this convergence is uniform over $x \in \bigcup_{k=1}^{\kappa} [a_k + \eta, b_k - \eta]$, for any small $\eta > 0$. Finally, for any small $\eta > 0$, there exists a finite nonrandom constant $\widetilde{K}$ such that almost surely, over the set $x \in \bigcup_{k=1}^{\kappa} [a_k - \eta, b_k + \eta]$, $|\widetilde{\varphi}_n(x)|$ is uniformly bounded by $\widetilde{K}$, for all n large enough.*

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 Ledoit and Wolf (2015, 2017) 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 3.2 (QuEST). *For any given n and p, $Q_{n,p}$ maps $\mathbf{t} := (t_1, \ldots, t_p) \in [0, +\infty)^p$ into*

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

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

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:

$$m = \frac{1}{p} \sum_{j=1}^{p} \frac{1}{t_j \left(1 - \frac{p}{n} - \frac{p}{n}\, x\, m\right) - x} \ . \tag{3.2}$$

Theorem 3.1 (Ledoit and Wolf (2015)). *Suppose Assumptions 2–4 are satisfied. Define*

$$\widehat{\boldsymbol{\tau}}_n := \underset{\mathbf{t} \in (0, +\infty)^p}{\operatorname{argmin}} \frac{1}{p} \sum_{i=1}^{p} \left[q^i_{n,p}(\mathbf{t}) - \lambda_{n,i}\right]^2 , \tag{3.3}$$

where $Q_{n,p}(\mathbf{t})$ is the QuEST function from Definition 3.2; 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 $\boldsymbol{\tau}_n := (\tau_{n,1}, \ldots, \tau_{n,p})'$ denote the population covariance matrix eigenvalues sorted in nondecreasing order w.l.o.g. Then

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

The function $\breve{m}^{\widehat{\boldsymbol{\tau}}_n}_{n,p}$ featured in the approximation (2.4) is a by-product of the QuEST function constructed by combining Equations (3.2)–(3.3). It estimates the complex-valued deterministic function of real argument $\breve{m}_{\mathrm{F}}$.

3.3 Dot Product of Population Eigenvalues with Sample Eigenvalues

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

$$\forall x, t \in \mathbb{R} \qquad \Theta_n(x, t) := \frac{1}{p} \sum_{i=1}^{p} \sum_{j=1}^{p} (u'_{n,i} v_{n,j})^2\, \mathbb{1}_{[\lambda_{n,i}, +\infty)}(x) \cdot \mathbb{1}_{[\tau_{n,j}, +\infty)}(t) \tag{3.4}$$

first introduced in Equation (6) of Ledoit and Péché (2011) 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 $u'_{n,i} v_{n,j}$would be something worth looking at. However, the sign is irrelevant, so we focus on the square $(u'_{n,i} v_{n,j})^2$ 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, $(u'_{n,i} v_{n,j})^2$ vanishes at the speed $1/p$, as can be seen from the following identities:

$$\frac{1}{p^2} \sum_{i=1}^{p} \sum_{j=1}^{p} (u'_{n,i} v_{n,j})^2 = \frac{1}{p^2} \sum_{i=1}^{p} u'_{n,i} \left(\sum_{j=1}^{p} v_{n,j} v'_{n,j} \right) u_{n,i} = \frac{1}{p^2} \sum_{i=1}^{p} u'_{n,i} u_{n,i} = \frac{1}{p} \ , \tag{3.5}$$

so it is more convenient to study $p(u'_{n,i}v_{n,j})^2$ instead. The average of the quantities of interest $p(u'_{n,i}v_{n,j})^2$ 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

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

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 Equation (3.6) delivers.

Theorem 3.2 (Ledoit and Péché (2011)). *Under Assumptions 2–4, $\forall \lambda, \tau \in \mathbb{R}$, $\Theta_n(\lambda,\tau)$ converges almost surely to some nonrandom bivariate c.d.f. $\Theta(\lambda,\tau) := \int_{-\infty}^{\lambda}\int_{-\infty}^{\tau} \theta(x,t)\,dH(t)\,dF(x)$, where*

$$\forall x, t \in \mathbb{R} \qquad \theta(x,t) := \frac{cxt}{\Big|t\big[1 - c - c\,x\,\breve{m}_F(x)\big] - x\Big|^2} \,. \tag{3.6}$$

The Radon-Nikodym derivative $\theta(\lambda_{n,i}, \tau_{n,j})$ is 'essentially like' the squared dot product $p(u'_{n,i}v_{n,j})^2$ for large p and n. In order to operationalize Equation (3.6), we need *bona fide* estimators for its ingredients, and they are provided by Section 3.2's QuEST function:

$$\widehat{\theta}_n(\lambda_{n,i}, \widehat{\tau}_{n,j}) := \frac{\frac{p}{n}\lambda_{n,i}\widehat{\tau}_{n,j}}{\Big|\widehat{\tau}_{n,j}\big[1 - \frac{p}{n} - \frac{p}{n}\,\lambda_{n,i}\,\breve{m}^{\widehat{\boldsymbol{\tau}}_n}_{n,p}(\lambda_{n,i})\big] - \lambda_{n,i}\Big|^2} \,. \tag{3.7}$$

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 Equations (3.1), (3.2), (3.6), and (3.7) appear to be descendents from each other, *it is because they are*. A graphical illustration in the case where the population eigenvalues are evenly spread in the interval $[1, 5]$ is given by Figure 3.1.

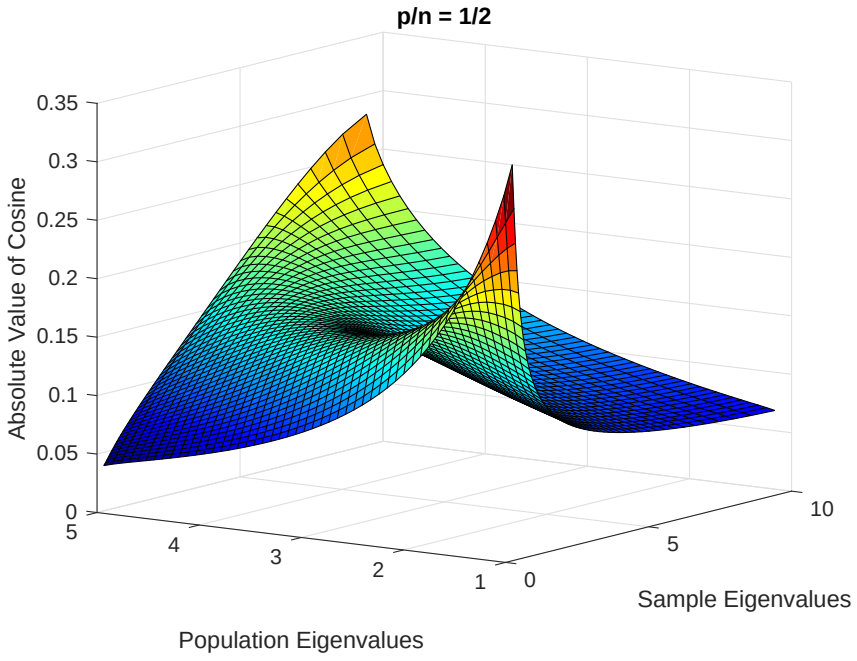

Figure 3.1: Absolute value of the cosine of the angle between population and sample eigenvectors. On the horizontal axes, eigenvectors are indexed by their respective eigenvalues.

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 3.2.

4 Asymptotically Optimal Nonlinear Shrinkage Estimators

The nonlinear shrinkage estimator called $\widehat{S}_n^*$ in Ledoit and Wolf (2018b) is optimal with respect to the Weighted Frobenius loss under large-dimensional asymptotics. 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 Ledoit and Wolf (2018b).

4.1 Four Specific Loss Functions

All remaining theorems are proven in Appendix C. We start with asymptotically optimal *bona fide* estimators based on Table 2.3.

Theorem 4.1 (Log-Euclidian)**.** *For any estimator $\widetilde{S}_n$ in Definition 3.1, the Log-Euclidian loss*

$$\mathcal{L}_n^{LE}(\Sigma_n, \widetilde{S}_n) := \frac{1}{p}\mathsf{Tr}\left[\left\{\log\left(\Sigma_n\right) - \log\left(\widetilde{S}_n\right)\right\}^2\right] , \tag{4.1}$$

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

$$\widehat{\varphi}_n^{LE}(\lambda_{n,i}) := \exp\left(\frac{1}{p}\sum_{j=1}^{p}\log(\widehat{\tau}_{n,j})\cdot\widehat{\theta}_n(\lambda_{n,i},\widehat{\tau}_{n,j})\right) , \tag{4.2}$$

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 3.1, 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 Equation (3.7). The resulting covariance matrix estimator is $\widehat{S}_n^{LE} := \sum_{i=1}^p \widehat{\varphi}_n^{LE}\left(\lambda_{n,i}\right)\cdot u_{n,i}u_{n,i}'$.

Theorem 4.2 (Fréchet)**.** *The Fréchet loss $\mathcal{L}_n^{FR\acute{E}}(\Sigma_n, \widetilde{S}_n) := \left\|\Sigma_n^{1/2} - \widetilde{S}_n^{1/2}\right\|_F^2/p$ converges almost surely to a deterministic limit that is minimized if $\widetilde{\varphi}_n(\lambda_{n,i})$ is equal to*

$$\widehat{\varphi}_n^{FR\acute{E}}(\lambda_{n,i}) := \left(\frac{1}{p}\sum_{j=1}^{p}\sqrt{\widehat{\tau}_{n,j}}\cdot\widehat{\theta}_n(\lambda_{n,i},\widehat{\tau}_{n,j})\right)^2 . \tag{4.3}$$

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

Theorem 4.3 (Quadratic)**.** *The Quadratic loss* $\mathcal{L}^{Q}\big(\Sigma,\widetilde{S}\big) \coloneqq \big\|\Sigma_n^{-1}\widetilde{S}_n - \mathbb{I}\big\|_F^2/p$ *converges almost surely to a deterministic limit that is minimized if* $\widetilde{\varphi}_n(\lambda_{n,i})$ *is equal to*

$$\widehat{\varphi}_n^{Q}(\lambda_{n,i}) \coloneqq \frac{\frac{1}{p}\sum_{j=1}^{p}\frac{1}{\widehat{\tau}_{n,j}}\cdot\widehat{\theta}_n(\lambda_{n,i},\widehat{\tau}_{n,j})}{\frac{1}{p}\sum_{j=1}^{p}\frac{1}{\widehat{\tau}_{n,j}^2}\cdot\widehat{\theta}_n(\lambda_{n,i},\widehat{\tau}_{n,j})}\ . \tag{4.4}$$

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

Theorem 4.4 (Inverse Quadratic)**.** *The Inverse Quadratic loss function, which is defined as* $\mathcal{L}^{QINV}\big(\Sigma,\widetilde{S}\big) \coloneqq \big\|\widetilde{S}^{-1}\Sigma - \mathbb{I}\big\|_F^2/p$*, converges almost surely to a deterministic limit minimized by*

$$\widehat{\varphi}_n^{QINV}(\lambda_{n,i}) \coloneqq \frac{\frac{1}{p}\sum_{j=1}^{p}\widehat{\tau}_{n,j}^2\cdot\widehat{\theta}_n(\lambda_{n,i},\widehat{\tau}_{n,j})}{\frac{1}{p}\sum_{j=1}^{p}\widehat{\tau}_{n,j}\cdot\widehat{\theta}_n(\lambda_{n,i},\widehat{\tau}_{n,j})}\ . \tag{4.5}$$

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

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 uncountably infinite family of loss functions is what we call Generalized Frobenius.

Theorem 4.5 (Generalized Frobenius)**.** *For any invertible and continously differentiable function* γ *defined on* $(0,+\infty)$*, the Generalized Frobenius loss* $\mathcal{L}_n^{\gamma,F}(\Sigma_n,\widetilde{S}_n) \coloneqq \Big\|\gamma\big(\Sigma_n\big) - \gamma\big(\widetilde{S}_n\big)\Big\|_F^2/p$ *converges almost surely to a deterministic limit that is minimized if* $\widetilde{\varphi}_n(\lambda_{n,i})$ *is equal to*

$$\widehat{\varphi}_n^{\gamma}(\lambda_{n,i}) \coloneqq \gamma^{-1}\left(\frac{1}{p}\sum_{j=1}^{p}\gamma\left(\widehat{\tau}_{n,j}\right)\cdot\widehat{\theta}_n(\lambda_{n,i},\widehat{\tau}_{n,j})\right)\ . \tag{4.6}$$

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

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 and Leibler (1951) divergence. 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:

$$D_{KL}\big(\mathcal{N}(0, A_1) \| \mathcal{N}(0, A_2)\big) := \frac{1}{2p}\left\{\mathsf{Tr}\big[A_2^{-1}A_1\big] - \log\big[\det(A_2^{-1}A_1)\big] - p\right\} \ . \tag{4.7}$$

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 4.6 (Generalized Kullback-Leibler)**.** *For any invertible and continously differentiable function γ defined on $(0, +\infty)$, the Generalized Kullback-Leibler loss function*

$$\mathcal{L}_n^{\gamma,KL}(\Sigma_n, \widetilde{S}_n) := \frac{1}{2p}\left\{\mathsf{Tr}\left[\gamma\big(\widetilde{S}_n\big)^{-1}\gamma\big(\Sigma_n\big)\right] - \log\det\left[\gamma(\widetilde{S}_n)^{-1}\gamma(\Sigma_n)\right] - p\right\} \tag{4.8}$$

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 Equation (4.6) (for $i = 1, \ldots, p$).

Both uncountably infinite families of loss functions confirm the asymptotic optimality of the *same* uncountably 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, and 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 4.1. The three other nonlinear shrinkage formulas that do not fit into the mold of Equation (4.6) are just elementary combinations of $\widehat{\varphi}_n^{\gamma}(\cdot)$ for two different γ functions. ■

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; therefore, it is only positive *semi*-definite. There then exist some linear combinations of the original variables that falsely appear to have zero variance when you only look in-sample. In a sense, the sample covariance matrix, with its $p(p+1)/2$ degrees of freedom, 'overfits' the data set of dimension $n \times p$.

5.1 Finite-Sample Analysis

With respect to the loss functions studied in this paper, the optimal nonlinear shrinkage formula applied to the n non-zero 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 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 and present a counterpart to Tables 2.1–2.3, listing how to optimally shrink null sample eigenvalues.

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

Table 5.1: Formulas for shrinking null eigenvalues.

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

5.2 Large-Dimensional Asymptotic Framework

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:

$$\forall x \in \mathbb{R} \qquad \underline{F}_n(x) \xrightarrow{\text{a.s.}} \underline{F}(x) := (1-c)\mathbb{1}_{[0,+\infty)}(x) + cF(x) \; . \tag{5.1}$$

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 Equation (2.4) that pertains to the shrinkage of the null sample eigenvalues. It comes from Theorem 9 of Ledoit and Péché (2011):

$$\boxed{\frac{1}{p-n} \sum_{i=1}^{p-n+1} u_{n,i}' \gamma(\Sigma_n) u_{n,i} \approx \frac{1}{p} \sum_{j=1}^{p} \gamma(\widehat{\tau}_{n,j}) \cdot \frac{1}{\left(1 - \frac{n}{p}\right) \left[1 + \underline{\breve{m}}_{n,p}^{\widehat{\boldsymbol{\tau}}}(0) \widehat{\tau}_{n,j}\right]}} \; , \tag{5.2}$$

where $\widehat{\boldsymbol{\tau}}_n := (\widehat{\tau}_{n,j})_{j=1,\ldots,p}$ is, as before, the estimator of population eigenvalues obtained by numerically inverting the QuEST function, and $\underline{\breve{m}}_{n,p}^{\widehat{\boldsymbol{\tau}}}(0)$ is a strongly consistent estimator of $\breve{m}_{\underline{F}}(0)$ that is another by-product of the QuEST function (when $p > n$). As per Ledoit and Wolf (2015, Section 3.2.2), $m\underline{\breve{m}}_{n,p}^{\widehat{\boldsymbol{\tau}}}(0)$ is the unique solution $m \in (0, \infty)$ to the equation

$$\frac{1}{m} = \frac{1}{n} \sum_{j=1}^{p} \frac{\widehat{\tau}_{n,j}}{1 + \widehat{\tau}_{n,j} \, m} \; . \tag{5.3}$$

Equation (5.2) enables us to extend the squared-dot-product function $\theta(x,t)$ presented in Section 3.3 to handle $x = 0$. The next figure graphs

$$\theta(0,t) := \frac{1}{\left(1 - \frac{1}{c}\right) \left[1 + \breve{m}_{\underline{F}}(0) \, t\right]} \tag{5.4}$$

as a function of t for various values of the concentration ratio p/n. We use the same baseline scenario as in Figure 3.1: the population eigenvalues are evenly spread in the interval [1,5].

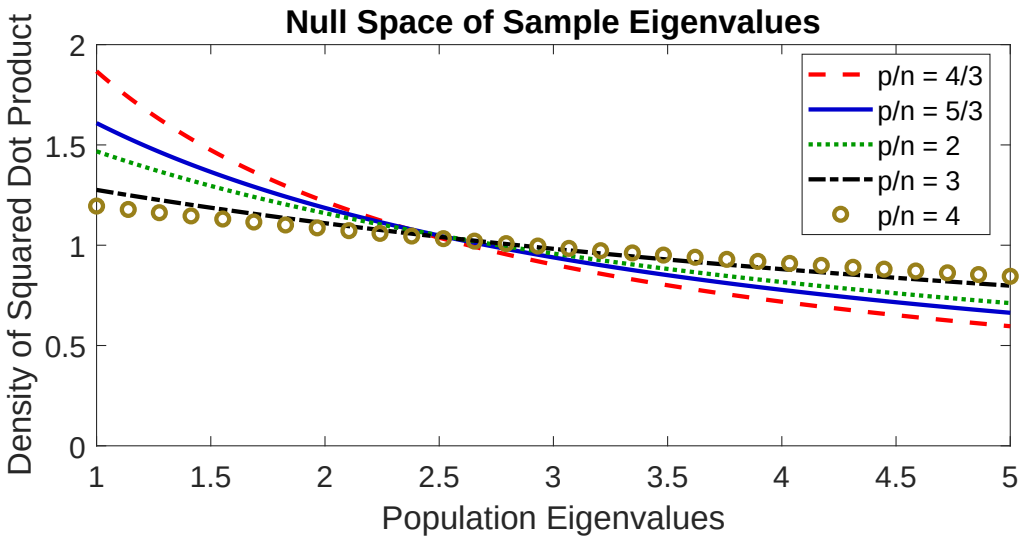

Figure 5.1: The Radon-Nykodym derivative $\theta(0,t)$ as a function of the population eigenvalues. This plot shows how aligned the null-space sample eigenvectors are with the population eigenvectors.

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 high p/n disorients the sample eigenvectors. The overall pattern is highly nonlinear, and could only be pinned down through Equations (5.3)–(5.4) 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 4.1–4.6 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 ($i = 1, \ldots, p-n$) is

$$\forall j = 1, \ldots, p \qquad \widehat{\theta}_n(\lambda_{n,i}, \widehat{\tau}_{n,j}) = \widehat{\theta}_n(0, \widehat{\tau}_{n,j}) := \frac{1}{\left(1 - \frac{n}{p}\right)\left[1 + \underline{\breve{m}}^{\widehat{\tau}}_{n,p}(0)\widehat{\tau}_{n,j}\right]} \,. \tag{5.5}$$

In order to show how this works, we need only state and prove the singular-case counterpart of Theorem 4.5, as the other theorems are adapted from $p < n$ to the $p > n$ case in similar fashion.

Theorem 5.1. *Under Assumptions 3–6, the Generalized Frobenius loss admits an almost sure (deterministic) limit, which is minimized by the nonlinear shrinkage formula*

$$\widehat{\varphi}^{\gamma}_n(\lambda_{n,i}) := \gamma^{-1}\left(\frac{1}{p}\sum_{j=1}^{p} \gamma\left(\widehat{\tau}_{n,j}\right) \cdot \widehat{\theta}_n(\lambda_{n,i}, \widehat{\tau}_{n,j})\right) \,, \tag{5.6}$$

where the bivariate function $\widehat{\theta}_n(x,t)$ *is given by Equation* (3.7) *for* $x > 0$*, and* $\widehat{\theta}_n(0,t)$ *is given by Equation* (5.5)*. The resulting covariance matrix estimator is* $\widehat{S}_n^{\gamma} = \sum_{i=1}^{p} \widehat{\varphi}_n^{\gamma}(\lambda_{n,i}) \cdot u_{n,i} u_{n,i}'$.

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 Bai and Silverstein (1998). We use the 12 loss functions from Tables 2.1–2.3. 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 Ledoit and Wolf (2018b): $\widehat{S}_n^{\circ}$ 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 Ledoit and Wolf (2004) are also computed for reference purposes. The results are averaged over 1,000 simulations.

6.2 Nonsingular Case

To produce the results of Table 6.1, matrix dimension is $p = 100$ and sample size is $n = 200$.

In each row, the performance of the best *bona fide* estimator is printed in bold. One can see that it is always the estimator tailor-made for the loss function of the row that wins. 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. The results for the identity matrix reported here are meaningful because that

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	5.755	14.644	14.771	7.382	**5.925**	7.747	6.441	6.297	6.016	16.094	8.226
Inverse Stein	0.152	0.326	0.710	0.184	**0.157**	0.216	0.171	0.174	0.161	0.464	0.226
Minimum Variance	1.095	2.721	2.757	1.370	**1.138**	1.162	1.144	1.163	1.148	1.172	1.359
Stein	0.150	0.690	0.310	0.289	0.213	**0.154**	0.168	0.168	0.186	0.222	0.513
Inverse Frobenius	0.048	0.144	0.852	0.098	0.069	**0.051**	0.055	0.054	0.060	0.069	0.126
Symmetrized Stein	0.329	1.016	1.020	0.473	0.370	0.371	**0.339**	0.342	0.347	0.686	0.739
Weighted Frobenius	0.228	1.016	0.504	0.377	0.317	**0.233**	0.251	0.256	0.281	0.336	0.743
Disutility	0.290	0.504	5.257	0.342	**0.298**	0.442	0.329	0.343	0.311	0.919	0.405
Log-Euclidian	0.291	0.859	0.756	0.427	0.329	0.324	0.301	**0.300**	0.307	0.598	0.637
Fréchet	0.286	0.772	0.585	0.367	0.300	0.347	0.302	0.299	**0.294**	0.703	0.504
Quadratic	0.292	4.212	1.013	1.289	0.978	0.462	0.647	0.668	0.803	**0.298**	2.927
Inverse Quadratic	0.260	0.503	9.490	0.376	0.449	1.104	0.685	0.737	0.576	2.642	**0.264**

Table 6.1: Average losses computed for various estimators when $p = 100$ and $n = 200$. Best numbers are in **bold face**.

is what we would get if we used the approach of Donoho et al. (2018), in the absence of outlying eigenvalues.

6.3 Comparison of Shrinkage Formulas

Confirming the ordering of Proposition 2.1, Figure 6.1 gives further insight into the loss functions by showing how the 7 estimators shrink the sample eigenvalues in this case.

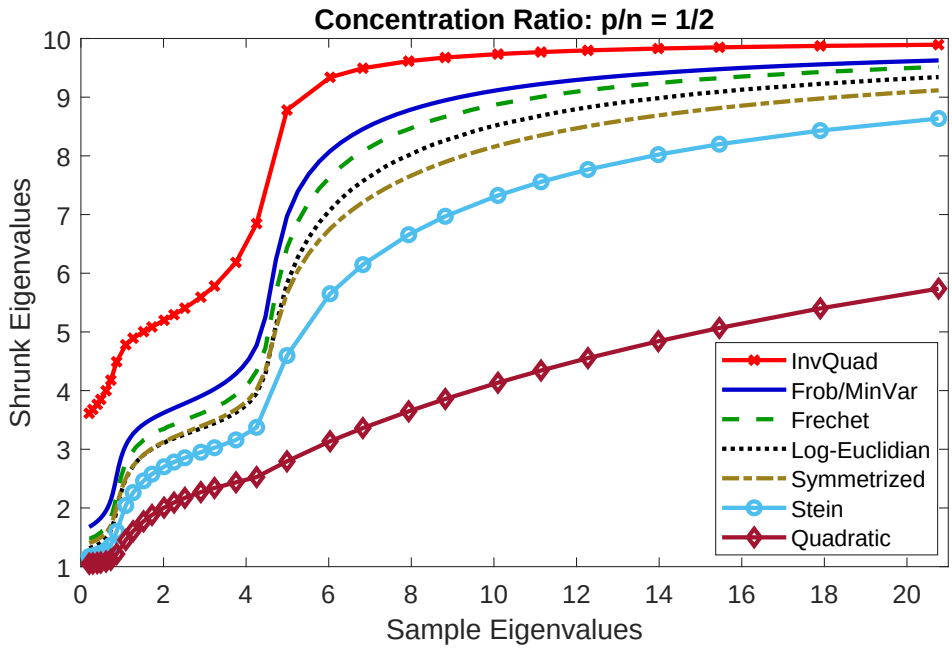

Figure 6.1: Comparison of 7 nonlinear shrinkage formulas.

The Quadratic and the Inverse Quadratic shrinkage formulas stand out as 'outliers', as shown by Proposition 2.1. In Table 6.1, 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 an excellent 'neutral' compromise that has strong foundations in the differential geometry of the manifold of tensors (a.k.a. positive definite matrices).

6.4 Singular Case

Table 6.2 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

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	11.250	14.644	11.774	**11.360**	15.343	12.590	12.559	11.688	22.044	17.560
Inverse Stein	0.274	0.326	0.280	**0.275**	0.418	0.308	0.315	0.285	0.729	0.358
Minimum Variance	2.221	2.721	2.271	**2.232**	2.255	2.239	2.253	2.241	2.301	2.362
Stein	0.290	0.690	0.510	0.496	**0.299**	0.356	0.347	0.407	0.339	1.071
Inverse Frobenius	0.091	0.144	0.128	0.126	**0.094**	0.107	0.104	0.114	0.102	0.163
Symmetrized Stein	0.656	1.016	0.789	0.772	0.716	0.665	**0.662**	0.693	1.068	1.428
Weighted Frobenius	0.397	1.015	0.707	0.697	**0.406**	0.475	0.470	0.557	0.463	1.933
Disutility	0.453	0.504	0.459	**0.455**	0.691	0.501	0.516	0.469	1.292	0.530
Log-Euclidian	0.587	0.859	0.687	0.672	0.636	0.595	**0.592**	0.614	0.914	1.123
Frechet	0.572	0.772	0.610	0.595	0.697	0.592	0.593	**0.577**	1.032	1.012
Quadratic	0.395	4.210	2.718	2.648	0.803	1.422	1.367	1.895	**0.490**	8.418
Inverse Quadratic	0.321	0.503	0.517	0.525	1.851	0.947	1.033	0.726	4.148	**0.322**

Table 6.2: Average losses computed for various estimators when $p = 200$ and $n = 100$. Best numbers are in **bold face**.

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

We examine two alternative approaches that make compromises in order to obtain formulas that are simpler than the ones developed in this paper. Linear shrinkage (Ledoit and Wolf, 2004) compromises by forcing all eigenvalues to be shrunk towards the same target with the same shrinkage intensity, and by considering only the Frobenius loss. The spiked-model approach (Donoho et al., 2018) compromises by assuming that the *bulk* of the population eigenvalues (meaning: all of them except for a vanishing fraction) are equal to each other.

In this Monte Carlo simulation, we take both the simpler linear shrinkage and the simpler spiked model 'outside of their comfort zone' by considering 8 different loss functions, and by considering specifications where the bulk of the population eigenvalues can be different from each other. 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 all of the ones in the intersection of the 12 that we consider in the present paper with the 18 for which Donoho et al. (2018) deliver 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. 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 one and setting the other half equal to $\overline{\tau} \in [1,5]$. It is only fair to allow bulk eigenvalues to not all be equal to each other: after all, this is the generic case, and the special case where all bulk eigenvalues are equal to each other is a measure-zero subset of the set of all possible eigenvalue combinations, so it is not necessarily representative of 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 $n = 202$.

Figure 6.2 displays the Percentage Relative Improvement in Average Loss (PRIAL):

$$\text{PRIAL}\left(\mathcal{L}^i, \widetilde{S}\right) := 100\% \cdot \left\{1 - \frac{\mathbb{E}\left[\mathcal{L}^i\left(\widehat{S}^{\text{FSOPT(i)}}, S\right)\right]}{\mathbb{E}\left[\mathcal{L}^i\left(\widehat{S}^{\text{FSOPT(i)}}, \widetilde{S}\right)\right]}\right\}, \tag{6.1}$$

where $\mathcal{L}^i$ denotes one of the eight loss functions listed above, $\widehat{S}^{\text{FSOPT(i)}}$ denotes the FSOPT estimator tailored to each specific loss function as per Tables 2.1–2.3, $\widetilde{S}$ is the estimator under consideration (whether linear shrinkage, spike shrinkage, or nonlinear shrinkage), and the expectation is approximated by the average of 1,000 Monte Carlo simulations. By construction,

the PRIAL of the sample covariance matrix is 0% whereas the PRIAL of the FSOPT estimator is 100%. The PRIAL measures how much of the potential for improvement relative to the sample covariance matrix is attained by a given estimator $\widetilde{S}$.

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 how spread out are the bulk population eigenvalues. Linear shrinkage has more of a mixed performance, but still manages to capture at least 50% of the potential improvement most of the time. It beats the sample covariance matrix with respect to all 8 loss functions, which shows that its attractiveness extends far beyond the Frobenius loss under which it was originally derived. It beats spike shrinkage as long as $\overline{\tau} \geq 2.5$ in all cases but one (the Quadratic loss, where they are essentially identical). Also worth noting is that linear shrinkage is the only estimator that keeps the same formula in all 8 subplots of Figure 6.2, so it is 'fighting with one hand tied behind the back' when it has to compete against the other two shrinkage estimators 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 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.

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.

7 Empirical Application

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

7.1 Data and Portfolio-Formation Rules

We download daily stock return data from the Center for Research in Security Prices (CRSP) starting on 01/01/1973 and ending on 12/31/2017. We restrict attention to stocks from the NYSE, AMEX, and NASDAQ stock exchanges.

For simplicity, we adopt the common convention that 21 consecutive trading days constitute

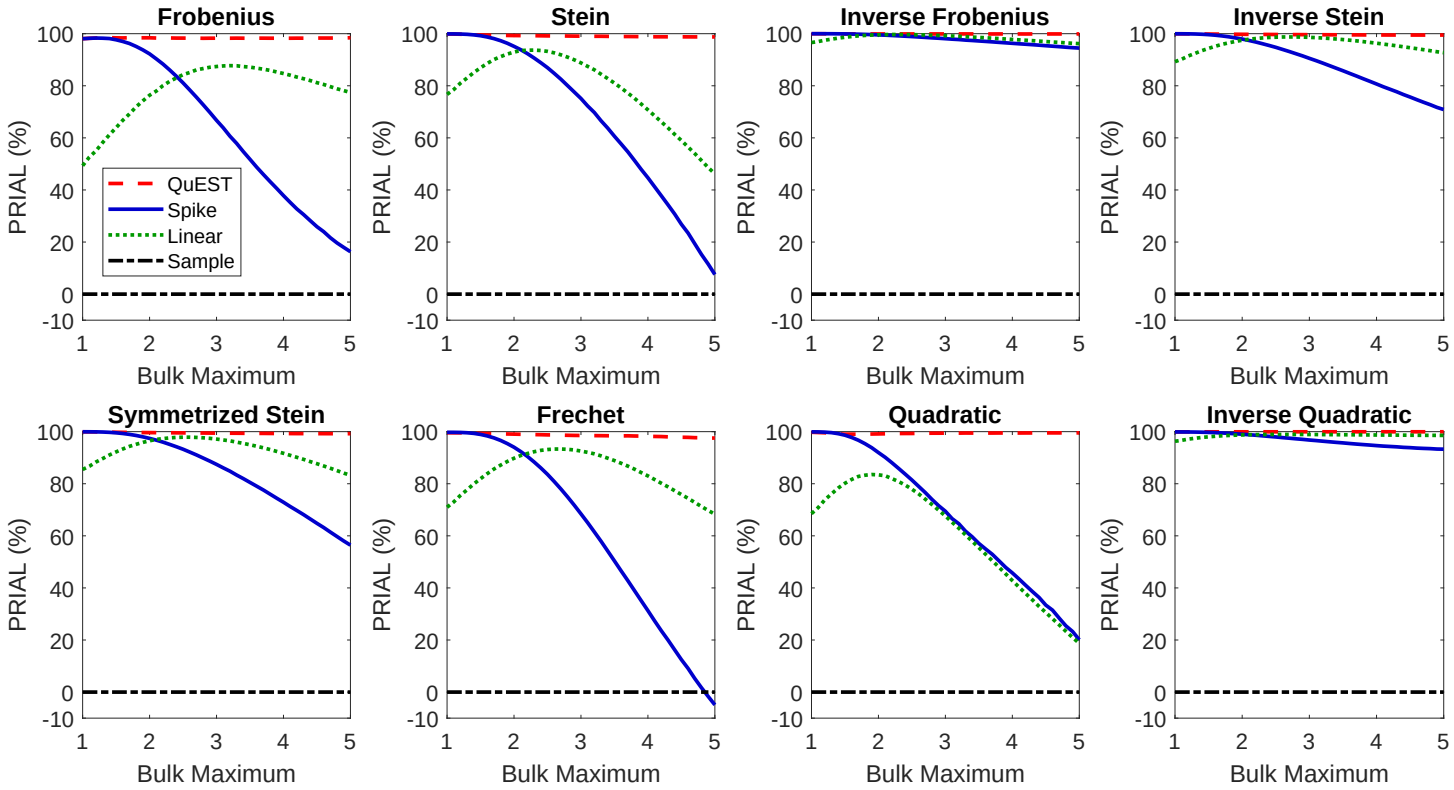

Figure 6.2: PRIAL as a function of the spread of the bulk of population eigenvalues for 4 *bona fide* estimators under 8 loss functions.

one 'month'. The out-of-sample period ranges from 01/16/1978 through 12/31/2017, resulting in a total of 480 months (or 10,080 days). All portfolios are updated monthly, which is standard practice in the finance literature. We denote the investment dates by $h = 1, \ldots, 480$. At any investment date h, a covariance matrix is estimated based on the most recent 1260 daily returns, which roughly corresponds to using five years of past data.

We consider the following portfolio sizes: $p \in \{100, 500, 1000\}$. For a given combination (h, p), the investment universe is obtained as follows. We find the set of stocks that have an almost complete return history over the most recent $n = 1260$ days as well as a complete return 'future' over the next 21 days. (The first restriction allows for up to 2.5% of missing returns over the most recent 1260 days, and replaces missing values by zero. The latter, 'forward-looking' restriction is not a feasible one in real life but is commonly applied in the finance literature.) We then look for possible pairs of highly correlated stocks, that is, pairs of stocks that have returns with a sample correlation exceeding 0.95 over the past 1260 days. In such pairs, if they should exist, we remove the stock with the lower market capitalization of the two on investment date h. (The reason is that we do not want to include highly similar stocks. In the early years, there are no such pairs; in the most recent years, there are never more than three such pairs.) Of the remaining set of stocks, we then pick the largest p stocks (as measured by their market capitalization on investment date h) as our investment universe. In this way, the investment universe changes relatively slowly from one investment date to the next.

7.2 Global Minimum Variance Portfolio

We consider the problem of estimating the global minimum variance (GMV) portfolio, in the absence of short-sales constraints. The problem goes back to Markowitz (1952) and is formulated as

$$\min_{w} w'\Sigma w \tag{7.1}$$

$$\text{subject to} \quad w'\mathbb{1} = 1 \ , \tag{7.2}$$

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

$$w = \frac{\Sigma^{-1}\mathbb{1}}{\mathbb{1}'\Sigma^{-1}\mathbb{1}} \ . \tag{7.3}$$

The natural strategy in practice is to replace the unknown Σ by an estimator $\widehat{\Sigma}$ in formula (7.3), yielding a feasible portfolio

$$\widehat{w} := \frac{\widehat{\Sigma}^{-1}\mathbb{1}}{\mathbb{1}'\widehat{\Sigma}^{-1}\mathbb{1}} \ . \tag{7.4}$$

Estimating the GMV portfolio is a 'clean' problem in terms of evaluating the quality of a covariance matrix estimator, since it abstracts from having to estimate the vector of expected returns at the same time. In addition, researchers have established that estimated GMV portfolios have desirable out-of-sample properties not only in terms of risk but also in terms of reward-to-risk (that is, in terms of the information ratio); for example, see Haugen and Baker

(1991), Jagannathan and Ma (2003), and Nielsen and Aylursubramanian (2008). As a result, such portfolios have become an addition to the large array of products sold by the mutual-fund industry. The following five portfolios are included in the study; they are all specific cases of portfolio (7.4), based on various covariance matrix estimators .

- **Identity:** the covariance matrix estimator $\widehat{\Sigma}$ is the identity matrix. This approach results in the equal-weighted portfolio that is quite popular in the finance literature and has been promoted by DeMiguel et al. (2009), among others.
- **Sample:** the covariance matrix estimator $\widehat{\Sigma}$ is the sample covariance matrix.
- **Lin:** the covariance matrix estimator $\widehat{\Sigma}$ is linear shrinkage.
- **Spike**: the covariance matrix estimator $\widehat{\Sigma}$ is spike shrinkage.
- **NonLin:** the covariance matrix estimator $\widehat{\Sigma}$ is nonlinear shrinkage.

All three shrinkage estimators use the Frobenius loss for two reasons. First, it can be considered a fair comparison, since linear shrinkage is only available for this loss. Second, the Frobenius loss leads to the same shrinkage formula as the Minimum Variance loss and the Disutility loss for nonlinear shrinkage and can thus be considered the 'right' loss function for portfolio selection problems. (Note that in all instances of 'overlap' of nonlinear and spike shrinkage, when two loss functions lead to the same shrinkage formula for nonlinear shrinkage, this is also true for spike shrinkage; hence, we can expect that the Frobenius loss also leads to the same shrinkage formula as the Minimum Variance loss for spike shrinkage, although the latter formula has not been established yet.)

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

- **AV:** We compute the average of the 10,080 out-of-sample log returns and then multiply by 252 to annualize.
- **SD:** We compute the standard deviation of the 10,080 out-of-sample log returns and then multiply by $\sqrt{252}$ to annualize.
- **IR:** We compute the (annualized) information ratio as the ratio AV/SD.

Our stance is that in the context of the GMV portfolio, the most important performance measure is the out-of-sample standard deviation, SD. The true (but unfeasible) GMV portfolio is given by (7.3). It is designed to minimize the variance (and thus the standard deviation) rather than to maximize the expected return or the information ratio. Therefore, any portfolio that implements the GMV portfolio should be primarily evaluated by how successfully it achieves this goal. A high out-of-sample average return, AV, and a high out-of-sample information ratio, IR, are naturally also desirable, but should be considered of secondary importance from the point of view of evaluating the quality of a covariance matrix estimator.

We also consider the question of whether one estimation model delivers a lower out-of-sample standard devation than another estimation model with statistical significance. Since we compare 5 estimation models, there are 10 pairwise comparisons. To avoid a multiple testing problem,

we restrict attention to the comparison between the two portfolios Spike and NonLin. For a given universe size, a two-sided p-value for the null hypothesis of equal standard deviations is obtained by the prewhitened HAC_{PW} method described in Ledoit and Wolf (2011, Section 3.1).

The results are presented in Table 7.1 and can be summarized as follows; unless stated otherwise, the findings are in terms of the standard deviation as performance measure.

- For all portfolio sizes p, NonLin is the best portfolio, Identity is the worst, and Sample is the second worst. The performances of Lin and Spike fall in between Sample and NonLin, with Lin being better than Spike for $p = 100$ and Spike being better than Lin for $p = 500, 1000$.
- Whereas Lin and NonLin are always better than Sample, this is not true for Spike: It is better than Sample for $p = 500, 1000$ but only equally as good for $p = 100$.
- For all portfolio sizes p, the outperformance of NonLin over Spike is significant at the 0.01 level.
- With respect to the information ratio, the ranking of the various portfolios is actually the same as with respect to the standard deviation. (Note that now larger numbers are better instead of smaller numbers.)
- The average by itself is rarely used as performance measure in the financial industry. Nevertheless, we can state that, on balance, Identity is best, Sample is worst, and there is no clear ranking of the three shrinkage portfolios.

Period: 01/16/1978–12/31/2017									
	$p = 100$			$p = 500$			$p = 1000$		
	AV	SD	IR	AV	SD	IR	AV	SD	IR
Structure-Free Models									
Identity	12.82	17.40	0.74	13.86	16.83	0.82	14.36	16.85	0.85
Sample	11.94	11.88	1.01	11.89	9.45	1.26	11.83	11.44	1.03
Lin	12.01	11.81	1.02	12.02	9.06	1.33	12.26	8.27	1.48
Spike	11.92	11.88	1.00	12.27	8.86	1.38	12.51	7.58	1.65
NL	11.94	**11.74*****	1.02	11.91	**8.63*****	1.38	12.28	**7.45*****	1.65

Table 7.1: Annualized performance measures (in percent) for various estimators of the GMV portfolio. AV stands for average; SD stands for standard deviation; and IR stands for information ratio. All measures are based on 10,080 daily out-of-sample returns from 01/16/1978 through 12/31/2017. In the columns labeled SD, the lowest number appears in **bold face**. In the row NonLin, significant outperformance over Spike in terms of SD is denoted by asterisks: *** denotes significance at the 0.01 level; ** denotes significance at the 0.05 level; and * denotes significance at the 0.1 level.

8 Conclusion

In this paper, we have

- developed an 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 our earlier work);
- 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 our earlier work);
- established an ordering of the nonlinear shrinkage formulas (from largest to smallest);
- delivered two infinite families of loss functions and their (correspondingly infinite family of) optimal nonlinear shrinkage formulas;
- and introduced a new loss function founded on the economic concept of utility maximization.

As a simpler alternative approach, Donoho et al. (2018) consider a *ménagerie* of 26 loss functions under the spiked covariance model of Johnstone (2001). 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. (2018) treat the spikes carefully, but they just collapse the bulk. This is a very restrictive policy that does not allow for any difference between population eigenvalues in the bulk. In the general case they are nonequal, so valuable information can be gleaned from the angle between sample and population eigenvectors, and from applying differentiated shrinkage inside the bulk. Other limitations are that these authors assume that the dimension is smaller than the sample size (which restricts applications); that there are no eigenvalues escaping the bulk from below; that the observations are Gaussian; and that there is just one bulk and not two or more bulks, which could happen due to the spectral separation phenomenon thoroughly documented by Bai and Silverstein (1998). We have demonstrated by both Monte Carlo simulations and an empirical application that applied researchers are well advised to ignore the KISS (*Keep it simple, statistician!*) principle and to upgrade instead from spike shrinkage to full-blown nonlinear shrinkage: They have, basically, nothing to lose but much to gain.

Having said this, Donoho et al. (2018) 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 7 different nonlinear shrinkage formulas, in addition to the 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, then certainly the Minimum Variance loss function is the appropriate one. An additional advantage is that a new technology has arisen for this purpose that is no more complex than kernel density estimation, and so is extremely fast and scalable to ultra-high dimensions (Ledoit and Wolf, 2018a). 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 those from Stein's loss.

References

Arsigny, V., Fillard, P., Pennec, X., and Ayache, N. (2006). Log-euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic Resonance in Medicine*, 56(2):411–421.

Bai, Z. D. and Silverstein, J. W. (1998). No eigenvalues outside the support of the limiting spectral distribution of large-dimensional random matrices. *Annals of Probability*, 26(1):316–345.

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Annals of Statistics*, 36(1):199–227.

DeMiguel, V., Garlappi, L., and Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the $1/N$ portfolio strategy? *Review of Financial Studies*, 22:1915–1953.

Donoho, D. L., Gavish, M., and Johnstone, I. M. (2018). Optimal shrinkage of eigenvalues in the spiked covariance model. *Annals of Statistics*, 46(4):1742–1778.

Dowson, D. and Landau, B. (1982). The Fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455.

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

Engle, R. F., Ledoit, O., and Wolf, M. (2019). Large dynamic covariance matrices. *Journal of Business & Economic Statistics*, 37(2):363–375.

Ghosh, M. and Sinha, B. (1987). Inadmissibility of the best equivariant estimators of the variance-covariance matrix, the precision matrix, and the generalized variance under entropy loss. *Statistics & Decisions*, 5(3-4):201–228.

Haff, L. (1979a). Estimation of the inverse covariance matrix: Random mixtures of the inverse wishart matrix and the identity. *Annals of Statistics*, pages 1264–1276.

Haff, L. (1979b). An identity for the Wishart distribution with applications. *Journal of Multivariate Analysis*, 9(4):531–544.

Haugen, R. A. and Baker, N. L. (1991). The efficient market inefficiency of capitalization-weighted stock portfolios. *Journal of Portfolio Management*, 17(3):35–40.

Jagannathan, R. and Ma, T. (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *Journal of Finance*, 54(4):1651–1684.

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

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

Jun, W. (2017). *On High-Dimensional Covariance Matrices Estimation*. PhD thesis, National University of Singapore, Department of Statistics and Applied Probability. Available online at `http://scholarbank.nus.edu.sg/handle/10635/135450`.

Kubokawa, T. and Konno, Y. (1990). Estimating the covariance matrix and the generalized variance under a symmetric loss. *Annals of the Institute of Statistical Mathematics*, 42(2):331–343.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Ledoit, O. and Péché, S. (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 150(1–2):233–264.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411.

Ledoit, O. and Wolf, M. (2011). Robust performance hypothesis testing with the variance. *Wilmott Magazine*, September:86–89.

Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Annals of Statistics*, 40(2):1024–1060.

Ledoit, O. and Wolf, M. (2015). Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions. *Journal of Multivariate Analysis*, 139(2):360–384.

Ledoit, O. and Wolf, M. (2017). Numerical implementation of the QuEST function. *Computational Statistics & Data Analysis*, 115:199–223.

Ledoit, O. and Wolf, M. (2018a). Analytical nonlinear shrinkage of large-dimensional covariance matrices. Working Paper ECON 264, Department of Economics, University of Zurich.

Ledoit, O. and Wolf, M. (2018b). Optimal estimation of a large-dimensional covariance matrix under Stein's loss. *Bernoulli*, 24(4B). 3791–3832.

Leung, P. L. and Muirhead, R. J. (1987). Estimation of parameter matrices and eigenvalues in manova and canonical correlation analysis. *The Annals of Statistics*, 15(4):1651–1666.

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

Nielsen, F. and Aylursubramanian, R. (2008). Far from the madding crowd — Volatility efficient indices. Research insights, MSCI Barra.

Selliah, J. B. (1964). *Estimation and Testing Problems in a Wishart Distribution*. PhD thesis, Stanford University, Department of Statistics.

Sharma, D. and Krishnamoorthy, K. (1985). Empirical Bayes estimators of normal covariance matrix. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 247–254.

Silverstein, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. *Journal of Multivariate Analysis*, 55:331–339.

Silverstein, J. W. and Choi, S. I. (1995). Analysis of the limiting spectral distribution of large-dimensional random matrices. *Journal of Multivariate Analysis*, 54:295–309.

Stein, C. (1956). 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*, pages 197–206. University of California Press.

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

Stein, C. (1986). Lectures on the theory of estimation of many parameters. *Journal of Mathematical Sciences*, 34(1):1373–1403.

Stein, C. M. (1969). Multivariate Analysis I. Technical Report No. 42, Department of Statistics, Stanford University. (Notes prepared by Morris L. Eaton.).

Stieltjes, T. J. (1894). Recherches sur les fractions continues. *Annales de la Faculté des Sciences de Toulouse 1^{re} Série*, 8(4):J1–J122.

Wigner, E. P. (1955). Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics*, 62(3):548–564.

A Portfolio Selection and the Disutility Loss

Here we explain how the Weighted Frobenius loss of Sharma and Krishnamoorthy (1985) applied to the precision matrix can be interpreted as quadratic *dis*utility. Consider the standard mean-variance optimization problem with quadratic utility function:

$$\max_w w'\mu - \frac{1}{2}\rho w'\Sigma w \ , \tag{A.1}$$

where μ denotes some vector of expected return selected by the end-user, and $\rho > 0$ the risk aversion parameter (cf. Markowitz (1952)). The first-order condition is $\mu - \rho\Sigma w = 0$, and the solution is $w = \Sigma^{-1}\mu/\rho$. In practice, we only observe an estimator $\widetilde{S}$ of the unobservable population covariance matrix Σ, so the plug-in estimator for the optimal weight vector is $\widetilde{w} = \widetilde{S}^{-1}\mu/\rho$. The quadratic utility associated with this vector is

$$\widetilde{w}'\mu - \frac{1}{2}\rho\widetilde{w}'\Sigma\widetilde{w} = \frac{1}{\rho}\mu'\widetilde{S}^{-1}\mu - \frac{1}{2\rho}\mu'\widetilde{S}^{-1}\Sigma\widetilde{S}^{-1}\mu \ . \tag{A.2}$$

At this point, the risk aversion coefficient ρ becomes irrelevant because, regardless of ρ, all investors want to find a covariance matrix estimator $\widetilde{S}$ that maximizes

$$\mu'\widetilde{S}^{-1}\mu - \frac{1}{2}\mu'\widetilde{S}^{-1}\Sigma\widetilde{S}^{-1}\mu \ . \tag{A.3}$$

As argued in further detail in Engle et al. (2019, Section 4.3), under large-dimensional asymptotics in conjunction with RMT, there is a key approximation:

$$\mu' A \mu \approx \|\mu\|^2 \mathsf{Tr}[A] \ . \tag{A.4}$$

From this we can streamline the objective function so as to make it equally adept at fitting the needs of all users who may have different views on the choice of vector μ:

$$\|\mu\|^2 \mathsf{Tr}\left[\widetilde{S}^{-1}\right] - \frac{1}{2}\|\mu\|^2 \mathsf{Tr}\left[\widetilde{S}^{-1}\Sigma\widetilde{S}^{-1}\right] . \tag{A.5}$$

The squared Euclidian norm of the linear constraint vector μ becomes irrelevant to the estimation process, so we are left with just maximizing $\mathsf{Tr}\big[\widetilde{S}^{-1}\big] - \frac{1}{2}\mathsf{Tr}\big[\widetilde{S}^{-1}\Sigma\widetilde{S}^{-1}\big]$.[1] This is obviously equivalent to *minimizing* with respect to the rotation-equivariant estimator $\widetilde{S}$ the shifted loss function

$$-\mathsf{Tr}\left[\widetilde{S}^{-1}\right] + \frac{1}{2}\mathsf{Tr}\left[\widetilde{S}^{-1}\Sigma\widetilde{S}^{-1}\right] + \frac{1}{2}\mathsf{Tr}\left[\Sigma^{-1}\right] = \frac{1}{2}\mathsf{Tr}\left[\left(\widetilde{S}^{-1} - \Sigma^{-1}\right)^2\Sigma\right] \ . \tag{A.6}$$

We recognize immediately the Weighted Frobenius loss function applied to the precision matrix

$$\mathcal{L}^{\mathrm{D}}\big(\Sigma, \widetilde{S}\big) \coloneqq \frac{\mathsf{Tr}\left[\left(\widetilde{S}^{-1} - \Sigma^{-1}\right)^2\Sigma\right]}{\mathsf{Tr}\big[\Sigma\big]} \ , \tag{A.7}$$

up to some multiplicative renormalizations. This approach nicely dovetails with the Minimum Variance loss function of Engle et al. (2019), as it gives the same optimal nonlinear shrinkage formula, but pins down the scaling factor internally rather than by appealing to the external argument of trace preservation.

[1] Note the close connection with the Minimum Variance loss function, which was essentially based on $\mathsf{Tr}\big[\widetilde{S}^{-1}\Sigma\widetilde{S}^{-1}\big]/\big(\mathsf{Tr}\big[\widetilde{S}^{-1}\big]\big)^2$. So, instead of dividing, we are subtracting here. Given that both of them are based on mean-variance portfolio optimization, it is reassuring to observe that they do not contradict each other.

B Finite-Sample Optimal Estimators for Various Losses

In this section, all loss functions are normalized by dimension so they admit an almost sure limit under large-dimensional asymptotics. Given that the objective is to optimize over the rotation-equivariant covariance matrix estimator $\widetilde{S}$, we call 'constant' any quantity that does not depend on $\widetilde{S}$.

B.1 Frobenius

$$\mathcal{L}^{\mathrm{F}}(\Sigma, \widetilde{S}) \coloneqq \frac{1}{p}\mathsf{Tr}\left[\left(\Sigma - \widetilde{S}\right)^2\right] = \frac{1}{p}\mathsf{Tr}\left[\left(\Sigma - U\widetilde{D}U'\right)\left(\Sigma - U\widetilde{D}U'\right)\right] \tag{B.1}$$

$$= \frac{1}{p}\mathsf{Tr}\left[U'\left(\Sigma - U\widetilde{D}U'\right)UU'\left(\Sigma - U\widetilde{D}U'\right)U\right] \tag{B.2}$$

$$= \frac{1}{p}\mathsf{Tr}\left[\left(U'\Sigma U - \widetilde{D}\right)^2\right] = \frac{1}{p}\sum_{i=1}^{p}\left(u_i'\Sigma u_i - \widetilde{d}_i\right)^2 + \text{constant}\ , \tag{B.3}$$

which is clearly minimized when $\widetilde{d}_i = u_i'\Sigma u_i$ for all $i = 1, \ldots, p$. ■

B.2 Inverse Stein

$$\mathcal{L}^{\mathrm{SINV}}(\Sigma, \widetilde{S}) \coloneqq \frac{1}{p}\mathsf{Tr}\left[\Sigma\widetilde{S}^{-1}\right] - \frac{1}{p}\log\left[\det\left(\Sigma\widetilde{S}^{-1}\right)\right] - 1 \tag{B.4}$$

$$= \frac{1}{p}\mathsf{Tr}\left[\Sigma U\widetilde{D}^{-1}U'\right] - \frac{1}{p}\log\left[\frac{\det(\Sigma)}{\det(\widetilde{S})}\right] - 1 \tag{B.5}$$

$$= \frac{1}{p}\mathsf{Tr}\left[U'\Sigma U\widetilde{D}^{-1}\right] - \frac{1}{p}\log\left[\det(\Sigma)\right] + \frac{1}{p}\log\left[\det(\widetilde{D})\right] - 1 \tag{B.6}$$

$$= \frac{1}{p}\sum_{i=1}^{p}\left[u_i'\Sigma u_i\widetilde{d}_i^{-1} + \log(\widetilde{d}_i)\right] + \text{constant} \tag{B.7}$$

$$\frac{\partial\mathcal{L}^{\mathrm{SINV}}(\Sigma, \widetilde{S})}{\partial\widetilde{d}_i} = -\frac{u_i'\Sigma u_i}{p\widetilde{d}_i^2} + \frac{1}{p\widetilde{d}_i} \qquad (\forall i = 1, \ldots, p) \tag{B.8}$$

$$\frac{\partial\mathcal{L}^{\mathrm{SINV}}(\Sigma, \widetilde{S})}{\partial\widetilde{d}_i} = 0 \Leftrightarrow \widetilde{d}_i = u_i'\Sigma u_i.\ \blacksquare \tag{B.9}$$

B.3 Minimum Variance

$$\mathcal{L}^{\mathrm{MV}}(\Sigma, \widetilde{S}) := \frac{\mathsf{Tr}\big[\widetilde{S}^{-1}\Sigma\widetilde{S}^{-1}\big]/p}{\Big(\mathsf{Tr}\big[\widetilde{S}^{-1}\big]/p\Big)^2} - \frac{1}{\mathsf{Tr}\big[\Sigma^{-1}\big]/p} \tag{B.10}$$

$$= p\frac{\mathsf{Tr}\big[U\widetilde{D}^{-1}U'\Sigma U\widetilde{D}^{-1}U'\big]}{\Big(\mathsf{Tr}\big[U\widetilde{D}^{-1}U'\big]\Big)^2} + \text{constant} \tag{B.11}$$

$$= p\frac{\mathsf{Tr}\big[\widetilde{D}^{-2}U'\Sigma U\big]}{\Big(\mathsf{Tr}\big[\widetilde{D}^{-1}\big]\Big)^2} + \text{constant} \tag{B.12}$$

$$= p\frac{\sum_{i=1}^p \widetilde{d}_i^{-2}u_i'\Sigma u_i}{\Big(\sum_{i=1}^p \widetilde{d}_i^{-1}\Big)^2} + \text{constant} \tag{B.13}$$

$$\frac{\partial\mathcal{L}^{\mathrm{MV}}(\Sigma, \widetilde{S})}{\partial\widetilde{d}_i} = p\frac{-2\widetilde{d}_i^{-3}u_i'\Sigma u_i\left(\sum_{j=1}^p \widetilde{d}_j^{-1}\right)^2 + 2\widetilde{d}_i^{-2}\left(\sum_{j=1}^p \widetilde{d}_j^{-1}\right)\left(\sum_{j=1}^p \widetilde{d}_j^{-2}u_j'\Sigma u_j\right)}{\left(\sum_{j=1}^p \widetilde{d}_j^{-1}\right)^4} \tag{B.14}$$

$$\frac{\partial\mathcal{L}^{\mathrm{MV}}(\Sigma, \widetilde{S})}{\partial\widetilde{d}_i} = 0 \Leftrightarrow \widetilde{d}_i^{-3}u_i'\Sigma u_i\left(\sum_{j=1}^p \widetilde{d}_j^{-1}\right)^2 = \widetilde{d}_i^{-2}\left(\sum_{j=1}^p \widetilde{d}_j^{-1}\right)\left(\sum_{j=1}^p \widetilde{d}_j^{-2}u_j'\Sigma u_j\right) \tag{B.15}$$

$$\Leftrightarrow \widetilde{d}_i = \text{scalar} \cdot u_i'\Sigma u_i \ , \tag{B.16}$$

where the scalar is independent of $i = 1, \ldots, p$. Any cursory inspection of the Minimum Variance loss function $\mathcal{L}^{\mathrm{MV}}(\Sigma, \widetilde{S})$ immediately reveals that the scalar cannot be determined internally, because multiplying the estimator $\widetilde{S}$ by any strictly positive scalar just washes out. Therefore, we have to invoke other arguments to make a choice. By preservation of the trace, the scalar should be set equal to one.

B.4 Stein

$$\mathcal{L}^{\mathrm{S}}(\Sigma, \widetilde{S}) := \frac{1}{p}\mathsf{Tr}\big[\Sigma^{-1}\widetilde{S}\big] - \frac{1}{p}\log\Big[\det\big(\Sigma^{-1}\widetilde{S}\big)\Big] - 1 \tag{B.17}$$

$$= \frac{1}{p}\mathsf{Tr}\big[\Sigma^{-1}U\widetilde{D}U'\big] - \frac{1}{p}\log\left[\frac{\det(\widetilde{S})}{\det(\Sigma)}\right] - 1 \tag{B.18}$$

$$= \frac{1}{p}\mathsf{Tr}\big[U'\Sigma^{-1}U\widetilde{D}\big] - \frac{1}{p}\log\big[\det(\widetilde{S})\big] + \frac{1}{p}\log\big[\det(\Sigma)\big] - 1 \tag{B.19}$$

$$= \frac{1}{p}\sum_{i=1}^{p}\Big[u_i'\Sigma^{-1}u_i\widetilde{d}_i - \log(\widetilde{d}_i)\Big] + \text{constant} \tag{B.20}$$

$$\frac{\partial\mathcal{L}^{\mathrm{S}}(\Sigma, \widetilde{S})}{\partial\widetilde{d}_i} = \frac{1}{p}u_i'\Sigma^{-1}u_i - \frac{1}{p\widetilde{d}_i} \qquad (\forall i = 1, \ldots, p) \tag{B.21}$$

$$\frac{\partial\mathcal{L}^{\mathrm{S}}(\Sigma, \widetilde{S})}{\partial\widetilde{d}_i} = 0 \Leftrightarrow \widetilde{d}_i = \frac{1}{u_i'\Sigma^{-1}u_i}. \blacksquare \tag{B.22}$$

B.5 Inverse Frobenius

$$\mathcal{L}^{\mathrm{FINV}}(\Sigma, \widetilde{S}) := \frac{1}{p}\mathsf{Tr}\Big[\big(\Sigma^{-1} - \widetilde{S}^{-1}\big)^2\Big] = \frac{1}{p}\mathsf{Tr}\Big[\big(\Sigma^{-1} - U\widetilde{D}^{-1}U'\big)\big(\Sigma^{-1} - U\widetilde{D}^{-1}U'\big)\Big] \tag{B.23}$$

$$= \frac{1}{p}\mathsf{Tr}\Big[U'\big(\Sigma^{-1} - U\widetilde{D}^{-1}U'\big)UU'\big(\Sigma^{-1} - U\widetilde{D}^{-1}U'\big)U\Big] \tag{B.24}$$

$$= \frac{1}{p}\mathsf{Tr}\Big[\big(U'\Sigma^{-1}U - \widetilde{D}^{-1}\big)^2\Big] = \frac{1}{p}\sum_{i=1}^{p}\left(u_i'\Sigma^{-1}u_i - \frac{1}{\widetilde{d}_i}\right)^2 + \text{constant}\ , \tag{B.25}$$

which is clearly minimized when $\widetilde{d}_i = \big(u_i'\Sigma^{-1}u_i\big)^{-1}$ for all $i = 1, \ldots, p$. $\blacksquare$

B.6 Symmetrized Stein

$$\mathcal{L}^{\mathrm{SSYM}}(\Sigma, \widetilde{S}) := \frac{1}{p}\mathsf{Tr}\Big[\Sigma^{-1}\widetilde{S} + \Sigma\widetilde{S}^{-1}\Big] - 2 = \frac{1}{p}\mathsf{Tr}\Big[\Sigma^{-1}U\widetilde{D}U' + \Sigma U\widetilde{D}^{-1}U'\Big] - 2 \tag{B.26}$$

$$= \frac{1}{p}\mathsf{Tr}\Big[U'\Sigma^{-1}U\widetilde{D} + U'\Sigma U\widetilde{D}^{-1}\Big] - 2 = \frac{1}{p}\sum_{i=1}^{p}\Big(u_i'\Sigma^{-1}u_i\widetilde{d}_i + u_i'\Sigma u_i\widetilde{d}_i^{-1}\Big) - 2 \tag{B.27}$$

$$\frac{\partial\mathcal{L}^{\mathrm{SSYM}}(\Sigma, \widetilde{S})}{\partial\widetilde{d}_i} = \frac{1}{p}u_i'\Sigma^{-1}u_i - \frac{u_i'\Sigma u_i}{p\widetilde{d}_i^{-2}} \tag{B.28}$$

$$\frac{\partial\mathcal{L}^{\mathrm{SSYM}}(\Sigma, \widetilde{S})}{\partial\widetilde{d}_i} = 0 \Leftrightarrow \widetilde{d}_i = \sqrt{\frac{u_i'\Sigma u_i}{u_i'\Sigma^{-1}u_i}}\ . \blacksquare \tag{B.29}$$

B.7 Weighted Frobenius

$$\mathcal{L}^{\mathrm{FW}}\big(\Sigma,\widetilde{S}\big) \coloneqq \frac{\mathsf{Tr}\left[\big(\widetilde{S}-\Sigma\big)^2\Sigma^{-1}\right]}{\mathsf{Tr}[\Sigma]} = \frac{\mathsf{Tr}\left[\big(U\widetilde{D}U'-\Sigma\big)^2\Sigma^{-1}\right]}{\mathsf{Tr}[\Sigma]} \tag{B.30}$$

$$= \frac{\mathsf{Tr}\left[U\widetilde{D}^2U'\Sigma^{-1}-2U\widetilde{D}U'+\Sigma\right]}{\mathsf{Tr}[\Sigma]} = \frac{\mathsf{Tr}\left[\widetilde{D}^2U'\Sigma^{-1}U'-2\widetilde{D}+\Sigma\right]}{\mathsf{Tr}[\Sigma]} \tag{B.31}$$

$$= \frac{\sum_{i=1}^p\left(\widetilde{d}_i^2u_i'\Sigma^{-1}u_i-2\widetilde{d}_i+\tau_i\right)}{\mathsf{Tr}[\Sigma]} \tag{B.32}$$

$$\frac{\partial\mathcal{L}^{\mathrm{FW}}\big(\Sigma,\widetilde{S}\big)}{\partial\widetilde{d}_i} = \frac{2\widetilde{d}_iu_i'\Sigma^{-1}u_i-2}{\mathsf{Tr}[\Sigma]} \tag{B.33}$$

$$\frac{\partial\mathcal{L}^{\mathrm{FW}}\big(\Sigma,\widetilde{S}\big)}{\partial\widetilde{d}_i} = 0 \Leftrightarrow \widetilde{d}_i = \frac{1}{u_i'\Sigma^{-1}u_i} \blacksquare \tag{B.34}$$

B.8 Disutility

$$\mathcal{L}^{\mathrm{D}}\big(\Sigma,\widetilde{S}\big) \coloneqq \frac{\mathsf{Tr}\left[\big(\widetilde{S}^{-1}-\Sigma^{-1}\big)^2\Sigma\right]}{\mathsf{Tr}[\Sigma^{-1}]} = \frac{\mathsf{Tr}\left[\big(U\widetilde{D}^{-1}U'-\Sigma^{-1}\big)^2\Sigma\right]}{\mathsf{Tr}[\Sigma^{-1}]} \tag{B.35}$$

$$= \frac{\mathsf{Tr}\left[U\widetilde{D}^{-2}U'\Sigma-2U\widetilde{D}^{-1}U'+\Sigma^{-1}\right]}{\mathsf{Tr}[\Sigma^{-1}]} = \frac{\mathsf{Tr}\left[\widetilde{D}^{-2}U'\Sigma U'-2\widetilde{D}^{-1}+\Sigma^{-1}\right]}{\mathsf{Tr}[\Sigma^{-1}]} \tag{B.36}$$

$$= \frac{\sum_{i=1}^p\left(\widetilde{d}_i^{-2}u_i'\Sigma u_i-2\widetilde{d}_i^{-1}+\tau_i^{-1}\right)}{\mathsf{Tr}[\Sigma^{-1}]} \tag{B.37}$$

$$\frac{\partial\mathcal{L}^{\mathrm{D}}\big(\Sigma,\widetilde{S}\big)}{\partial\widetilde{d}_i} = \frac{-2\widetilde{d}_i^{-3}u_i'\Sigma u_i+2\widetilde{d}_i^{-2}}{\mathsf{Tr}[\Sigma]} \tag{B.38}$$

$$\frac{\partial\mathcal{L}^{\mathrm{D}}\big(\Sigma,\widetilde{S}\big)}{\partial\widetilde{d}_i} = 0 \Leftrightarrow \widetilde{d}_i = u_i'\Sigma u_i \blacksquare \tag{B.39}$$

B.9 Log-Euclidian

$$\mathcal{L}^{\mathrm{LE}}(\Sigma,\widetilde{S}) \coloneqq \frac{1}{p}\mathsf{Tr}\left[\big\{\log(\Sigma)-\log(\widetilde{S})\big\}^2\right] = \frac{1}{p}\mathsf{Tr}\left[\big\{\log(\Sigma)-\log(U\widetilde{D}U')\big\}^2\right] \tag{B.40}$$

$$= \frac{1}{p}\mathsf{Tr}\left[\big\{\log(\Sigma)-U\log(\widetilde{D})U'\big\}^2\right] = \frac{1}{p}\mathsf{Tr}\left[\big\{U'\log(\Sigma)U-\log(\widetilde{D})\big\}^2\right] \tag{B.41}$$

$$= \frac{1}{p}\sum_{i=1}^p\big\{u_i'\log(\Sigma)u_i-\log(\widetilde{d}_i)\big\}^2 + \text{constant}\ , \tag{B.42}$$

which is clearly minimized when $\widetilde{d}_i = \exp\big[u_i'\log(\Sigma)u_i\big]$ for all $i=1,\ldots,p$. $\blacksquare$

B.10 Fréchet

$$\mathcal{L}^{\text{FRÉ}}\big(\Sigma,\widetilde{S}\big) \coloneqq \frac{1}{p}\mathsf{Tr}\left[\Sigma+\widetilde{S}-2\Sigma^{1/2}\widetilde{S}^{1/2}\right] = \frac{1}{p}\mathsf{Tr}\left[\Sigma+U\widetilde{D}U'-2\Sigma^{1/2}U\widetilde{D}^{1/2}U'\right] \tag{B.43}$$

$$= \frac{1}{p}\mathsf{Tr}\left[U'\Sigma U+\widetilde{D}-2U'\Sigma^{1/2}U\widetilde{D}^{1/2}\right] \tag{B.44}$$

$$= \frac{1}{p}\sum_{i=1}^{p}\left(\widetilde{d}_i-2u_i'\Sigma^{1/2}u_i\sqrt{\widetilde{d}_i}\right)+\text{constant} \tag{B.45}$$

$$\frac{\partial\mathcal{L}^{\text{FRÉ}}(\Sigma,\widetilde{S})}{\partial\widetilde{d}_i} = 1-\frac{u_i'\Sigma^{1/2}u_i}{\sqrt{\widetilde{d}_i}} \tag{B.46}$$

$$\frac{\partial\mathcal{L}^{\text{FRÉ}}(\Sigma,\widetilde{S})}{\partial\widetilde{d}_i} = 0 \Leftrightarrow \widetilde{d}_i = \left(u_i'\Sigma^{1/2}u_i\right)^2 \ . \blacksquare \tag{B.47}$$

B.11 Quadratic

$$\mathcal{L}^{\text{Q}}\big(\Sigma,\widetilde{S}\big) \coloneqq \frac{1}{p}\left\|\Sigma^{-1}\widetilde{S}-\mathbb{I}\right\|_{\text{F}}^2 = \frac{1}{p}\mathsf{Tr}\left[\left(\Sigma^{-1}\widetilde{S}-\mathbb{I}\right)\left(\Sigma^{-1}\widetilde{S}-\mathbb{I}\right)'\right] \tag{B.48}$$

$$= \frac{1}{p}\mathsf{Tr}\left[\left(\Sigma^{-1}\widetilde{S}-\mathbb{I}\right)\left(\widetilde{S}\Sigma^{-1}-\mathbb{I}\right)\right] = \frac{1}{p}\mathsf{Tr}\left[\left(\widetilde{S}\Sigma^{-1}-\mathbb{I}\right)\left(\Sigma^{-1}\widetilde{S}-\mathbb{I}\right)\right] \tag{B.49}$$

$$= \frac{1}{p}\mathsf{Tr}\left[\widetilde{S}\Sigma^{-2}\widetilde{S}-2\Sigma^{-1}\widetilde{S}+\mathbb{I}\right] \tag{B.50}$$

$$= \frac{1}{p}\mathsf{Tr}\left[U\widetilde{D}U'\Sigma^{-2}U\widetilde{D}U'-2\Sigma^{-1}U\widetilde{D}U'+\mathbb{I}\right] \tag{B.51}$$

$$= \frac{1}{p}\mathsf{Tr}\left[\widetilde{D}U'\Sigma^{-2}U\widetilde{D}-2U'\Sigma^{-1}U\widetilde{D}+\mathbb{I}\right] \tag{B.52}$$

$$= \frac{1}{p}\sum_{i=1}^{p}\widetilde{d}_i^2\cdot u_i'\Sigma^{-2}u_i-\frac{2}{p}\sum_{i=1}^{p}\widetilde{d}_i\cdot u_i'\Sigma^{-1}u_i+1 \tag{B.53}$$

The first-order condition (FOC) is obtained as follows:

$$\frac{\partial\mathcal{L}^{\text{Q}}\big(\Sigma,\widetilde{S}\big)}{\widetilde{d}_i} = \frac{2}{p}\widetilde{d}_i\cdot u_i'\Sigma^{-2}u_i-\frac{2}{p}u_i'\Sigma^{-1}u_i \tag{B.54}$$

$$\text{FOC: } \widetilde{d}_i\cdot u_i'\Sigma^{-2}u_i = u_i'\Sigma^{-1}u_i \Leftrightarrow \widetilde{d}_i = \frac{u_i'\Sigma^{-1}u_i}{u_i'\Sigma^{-2}u_i} \ . \blacksquare \tag{B.55}$$

B.12 Inverse Quadratic

The algebraic manipulations are the same as above, except that Σ^{-1} becomes Σ and $\widetilde{S}$ becomes $\widetilde{S}^{-1}$, so there is no point repeating the intermediary steps.

$$\mathcal{L}^{\text{QINV}}\big(\Sigma, \widetilde{S}\big) := \frac{1}{p}\left\|\widetilde{S}^{-1}\Sigma - \mathbb{I}\right\|_{\text{F}}^2 \tag{B.56}$$

$$= \frac{1}{p}\widetilde{d}_i^{-2} u_i'\Sigma^2 u_i - \frac{2}{p}\sum_{i=1}^{p}\widetilde{d}_i^{-1} u_i'\Sigma u_i + 1 \tag{B.57}$$

$$\frac{\partial \mathcal{L}^{\text{QINV}}\big(\Sigma, \widetilde{S}\big)}{\partial \widetilde{d}_i} = -\frac{2}{p}\widetilde{d}_i^{-3}\cdot u_i'\Sigma^2 u_i + \frac{2}{p}\widetilde{d}_i^{-2}\cdot u_i'\Sigma u_i \tag{B.58}$$

$$\text{FOC:}\quad \frac{\partial \mathcal{L}^{\text{QINV}}\big(\Sigma, \widetilde{S}\big)}{\partial \widetilde{d}_i} = 0 \Leftrightarrow \widetilde{d}_i = \frac{u_i'\Sigma^2 u_i}{u_i'\Sigma^2 u_i}\ . \blacksquare \tag{B.59}$$

B.13 Generalized Frobenius

To defuse a well-known source of confusion, we use the notation $\gamma^{-1}(x)$ to signify the inverse function of the invertible function γ, and $\gamma(x)^{-1}$ to signify one divided by $\gamma(x)$. For example, if $\gamma(x) = x^3$ then $\gamma^{-1}(x) = \sqrt[3]{x}$ and $\gamma(x)^{-1} = 1/x^3$.

$$\mathcal{L}^{\gamma,\text{F}}(\Sigma, \widetilde{S}) := \frac{1}{p}\mathsf{Tr}\left[\left\{\gamma(\Sigma) - \gamma(\widetilde{S})\right\}^2\right] = \frac{1}{p}\mathsf{Tr}\left[\left\{\gamma(\Sigma) - \gamma(U\widetilde{D}U')\right\}^2\right] \tag{B.60}$$

$$= \frac{1}{p}\mathsf{Tr}\left[\left\{\gamma(\Sigma) - U\gamma(\widetilde{D})U'\right\}^2\right] = \frac{1}{p}\mathsf{Tr}\left[\left\{U'\gamma(\Sigma)U - \gamma(\widetilde{D})\right\}^2\right] \tag{B.61}$$

$$= \frac{1}{p}\mathsf{Tr}\left[\left\{U'\gamma(\Sigma)U\right\}^2 - 2U'\gamma(\Sigma)U\gamma(\widetilde{D}) + \left\{\gamma(\widetilde{D})\right\}^2\right] \tag{B.62}$$

$$= \frac{1}{p}\mathsf{Tr}\left[\gamma(\Sigma)^2 - 2U'\gamma(\Sigma)U\gamma(\widetilde{D}) + \gamma(\widetilde{D})^2\right] \tag{B.63}$$

$$= \frac{1}{p}\sum_{i=1}^{p}\left\{\gamma(\tau_i)^2 - 2u_i'\gamma(\Sigma)u_i\gamma(\widetilde{d}_i) + \gamma(\widetilde{d}_i)^2\right\} \tag{B.64}$$

$$\frac{\partial \mathcal{L}^{\gamma,\text{F}}\big(\Sigma, \widetilde{S}\big)}{\partial \widetilde{d}_i} = -2u_i'\gamma(\Sigma)u_i\gamma'(\widetilde{d}_i) + 2\gamma'(\widetilde{d}_i)\gamma(\widetilde{d}_i) \qquad \forall i = 1, \ldots, p \tag{B.65}$$

$$\text{FOC:}\quad \frac{\partial \mathcal{L}^{\gamma,\text{F}}\big(\Sigma, \widetilde{S}\big)}{\partial \widetilde{d}_i} = 0 \Leftrightarrow \widetilde{d}_i = \gamma^{-1}\left(u_i'\gamma(\Sigma)u_i\right)\ . \blacksquare \tag{B.66}$$

B.14 Generalized Kullback-Leibler Divergence

$$\mathcal{L}^{\gamma,\mathrm{KL}}(\Sigma,\widetilde{S}) \coloneqq \frac{1}{2p}\left\{\mathsf{Tr}\left[\gamma(\widetilde{S})^{-1}\gamma(\Sigma)\right] - \log\det\left[\gamma(\widetilde{S})^{-1}\gamma(\Sigma)\right] - p\right\} \tag{B.67}$$

$$= \frac{1}{2p}\left\{\mathsf{Tr}\left[\gamma(U\widetilde{D}U')^{-1}\gamma(\Sigma)\right] - \log\det\left[\gamma(\widetilde{S})^{-1}\gamma(\Sigma)\right] - p\right\} \tag{B.68}$$

$$= \frac{1}{2p}\left\{\mathsf{Tr}\left[U\gamma(\widetilde{D})^{-1}U'\gamma(\Sigma)\right] - \log\det\left[\gamma(\widetilde{S})^{-1}\gamma(\Sigma)\right] - p\right\} \tag{B.69}$$

$$= \frac{1}{2p}\left\{\mathsf{Tr}\left[\gamma(\widetilde{D})^{-1}U'\gamma(\Sigma)U\right] + \log\left[\det\left(\gamma(\widetilde{S})\right)\right] - \log\left[\det\left(\gamma(\Sigma)\right)\right] - p\right\}$$

$$= \frac{1}{2p}\sum_{i=1}^{p}\left\{\frac{u_i'\gamma(\Sigma)u_i}{\gamma(\widetilde{\delta}_i)} + \log\left[\gamma(\widetilde{\delta}_i)\right] - \log\left[\gamma(\tau_i)\right] - 1\right\} \tag{B.70}$$

$$\frac{\partial\mathcal{L}^{\gamma,\mathrm{KL}}(\Sigma,\widetilde{S})}{\partial\widetilde{d}_i} = -\frac{u_i'\gamma(\Sigma)u_i}{\gamma(\widetilde{d}_i)^2}\gamma'(\widetilde{d}_i) + \frac{\gamma'(\widetilde{d}_i)}{\gamma(\widetilde{d}_i)} \qquad \forall i = 1,\ldots,p \tag{B.71}$$

$$\frac{\partial\mathcal{L}^{\gamma,\mathrm{KL}}(\Sigma,\widetilde{S})}{\partial\widetilde{d}_i} = 0 \Leftrightarrow \gamma(\widetilde{d}_i) = u_i'\gamma(\Sigma)u_i \Leftrightarrow \widetilde{d}_i = \gamma^{-1}\left(u_i'\gamma(\Sigma)u_i\right) \ . \blacksquare \tag{B.72}$$

C Proofs of Theorems in Sections 4 and 5

Theorems 4.1 and 4.2 are special cases of Theorem 4.5 with $\gamma(x) = \log(x)$, resp. $\gamma(x) = \sqrt{x}$.

C.1 Quadratic

Proposition C.1. *Under Assumptions 2–5,*

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

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 Appendix B.11 we have:

$$\mathcal{L}_n^{\mathrm{Q}}(\Sigma_n,\widetilde{S}_n) = \frac{1}{p}\sum_{i=1}^{p}\widetilde{d}_{n,i}^2\cdot u_{n,i}'\Sigma_n^{-2}u_{n,i} - \frac{2}{p}\sum_{i=1}^{p}\widetilde{d}_{n,i}\cdot u_{n,i}'\Sigma_n^{-1}u_{n,i} + 1 \tag{C.2}$$

$$= \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p}\left[\frac{\widetilde{d}_{n,i}^2}{\tau_{n,j}^2} - 2\frac{\widetilde{d}_{n,i}}{\tau_{n,j}} + 1\right]\cdot\left(u_i'v_j\right)^2 \tag{C.3}$$

$$= \int_a^b\int_{-\infty}^{+\infty}\left[\frac{\widetilde{\varphi}_n(\lambda_{n,i})^2}{\tau_{n,j}^2} - 2\frac{\widetilde{\varphi}_n(\lambda_{n,i})}{\tau_{n,j}} + 1\right]d^2\Theta_n(x,t)\ , \tag{C.4}$$

where Θ_n is the random bivariate function from Equation (3.4)). By applying the technique from the proof of Theorem 3.1 of Ledoit and Wolf (2018b), and by using Theorem 3.2 to handle the function Θ_n, it follows that

$$\mathcal{L}_n^{\mathrm{Q}}(\Sigma_n,\widetilde{S}_n) \xrightarrow{a.s.} \int_a^b\int_{-\infty}^{+\infty}\left[\frac{\widetilde{\varphi}(x)^2}{t^2} - 2\frac{\widetilde{\varphi}(x)}{t} + 1\right]\theta(x,t)\,dx\,dt\ , \tag{C.5}$$

where, as per Equation (3.6),

$$\forall x \in [a,b] \quad \forall t \in \mathbb{R} \qquad \theta(x,t) \coloneqq \frac{cxt}{\left|t\left[1-c-cx\breve{m}_{\mathrm{F}}(x)\right]-x\right|^2} \ . \ \blacksquare \tag{C.6}$$

Proposition C.1 lets us characterize the asymptotically optimal nonlinear shrinkage function under Quadratic loss.

Corollary C.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 (C.1) of the Quadratic loss if and only if its limiting shrinkage function $\widetilde{\varphi}$ verifies $\forall x \in \mathsf{Supp}(F)$, $\widetilde{\varphi}(x) = \widehat{\varphi}^{Q}(x)$, where*

$$\forall x \in \mathsf{Supp}(F) \qquad \widehat{\varphi}^{Q}(x) \coloneqq \frac{\displaystyle\int_{-\infty}^{+\infty} \frac{1}{t} \cdot \theta(x,t)\, dH(t)}{\displaystyle\int_{-\infty}^{+\infty} \frac{1}{t^2} \cdot \theta(x,t)\, dH(t)} \ . \tag{C.7}$$

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

$$\int_{-\infty}^{+\infty} \left[\frac{\widetilde{\varphi}(x)^2}{t^2} - 2\frac{\widetilde{\varphi}(x)}{t} + 1\right] \theta(x,t)\, dH(t) \ . \tag{C.8}$$

The partial derivative of (C.8) with respect to $\widetilde{\varphi}(x)$ is

$$\int_{-\infty}^{+\infty} \left[\frac{2\widetilde{\varphi}(x)}{t^2} - \frac{2}{t}\right] \theta(x,t)\, dH(t) \ . \tag{C.9}$$

The first-order condition is

$$\varphi(x) \int_{-\infty}^{+\infty} \frac{1}{t^2}\, \theta(x,t)\, dH(t) = \int_{-\infty}^{+\infty} \frac{1}{t}\, \theta(x,t)\, dH(t) \ . \tag{C.10}$$

The solution is

$$\varphi(x) = \frac{\displaystyle\int_{-\infty}^{+\infty} \frac{1}{t}\, \theta(x,t)\, dH(t)}{\displaystyle\int_{-\infty}^{+\infty} \frac{1}{t^2}\, \theta(x,t)\, dH(t)} \ . \ \blacksquare \tag{C.11}$$

The proof of Theorem 4.3 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 $\widehat{H}_n(t) \coloneqq \sum_{i=j}^{p} \mathbb{1}_{[\widehat{\tau}_{n,j},+\infty]}(t)/p$; and to the unobservable quantity $\theta(x)$ corresponds the plug-in estimator $\widehat{\theta}_n(x,t)$ from Equation (3.7). 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 Ledoit and Wolf's (2018b) Theorem 5.2. ■

C.2 Inverse Quadratic

Proposition C.2. *Under Assumptions 2–5,*

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

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

$$\mathcal{L}_n^{\text{QINV}}(\Sigma_n, \widetilde{S}_n) = \frac{1}{p}\sum_{i=1}^{p} \widetilde{d}_{n,i}^{-2} \cdot u_{n,i}' \Sigma_n^2 u_{n,i} - \frac{2}{p}\sum_{i=1}^{p} \widetilde{d}_{n,i}^{-1} \cdot u_{n,i}' \Sigma_n u_{n,i} + 1 \tag{C.13}$$

$$= \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p} \left[\frac{\tau_{n,j}^2}{\widetilde{d}_{n,i}^2} - 2\frac{\tau_{n,j}}{\widetilde{d}_{n,i}} + 1\right] \cdot (u_i' v_j)^2 \tag{C.14}$$

$$= \int_a^b \int_{-\infty}^{+\infty} \left[\frac{\tau_{n,j}^2}{\widetilde{\varphi}_n(\lambda_{n,i})^2} - 2\frac{\tau_{n,j}}{\widetilde{\varphi}_n(\lambda_{n,i})} + 1\right] d^2\Theta_n(x,t) \ . \tag{C.15}$$

By applying the technique from the proof of Theorem 3.1 of Ledoit and Wolf (2018b), and by using Theorem 3.2:

$$\mathcal{L}_n^{\text{QINV}}(\Sigma_n, \widetilde{S}_n) \xrightarrow{a.s.} \int_a^b \int_{-\infty}^{+\infty} \left[\frac{t^2}{\widetilde{\varphi}(x)^2} - 2\frac{t}{\widetilde{\varphi}(x)} + 1\right] \theta(x,t)\, dx\, dt \ . \blacksquare \tag{C.16}$$

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

$$\forall x \in \mathsf{Supp}(F) \qquad \widehat{\varphi}^{QINV}(x) := \frac{\displaystyle\int_{-\infty}^{+\infty} t^2 \cdot \theta(x,t)\, dH(t)}{\displaystyle\int_{-\infty}^{+\infty} t \cdot \theta(x,t)\, dH(t)} \ . \tag{C.17}$$

Proof. If we fix $x \in \mathsf{Supp}(F)$, 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

$$\int_{-\infty}^{+\infty} \left[\frac{t^2}{\widetilde{\varphi}(x)^2} - 2\frac{t}{\widetilde{\varphi}(x)} + 1\right] \theta(x,t)\, dH(t) \ . \tag{C.18}$$

The partial derivative of (C.18) with respect to $\widetilde{\varphi}(x)$ is

$$\int_{-\infty}^{+\infty} \left[-2\frac{t^2}{\widetilde{\varphi}(x)^3} + 2\frac{t}{\widetilde{\varphi}(x)^2}\right] \theta(x,t)\, dH(t) \ . \tag{C.19}$$

The first-order condition is

$$\varphi(x) \int_{-\infty}^{+\infty} t\, \theta(x,t)\, dH(t) = \int_{-\infty}^{+\infty} t^2\, \theta(x,t)\, dH(t) \ . \tag{C.20}$$

The solution is

$$\varphi(x) = \frac{\displaystyle\int_{-\infty}^{+\infty} t^2\, \theta(x,t)\, dH(t)}{\displaystyle\int_{-\infty}^{+\infty} t\, \theta(x,t)\, dH(t)} \ . \blacksquare \tag{C.21}$$

The proof of Theorem 4.4 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 Equation (3.7). 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 Ledoit and Wolf's (2018b) Theorem 5.2. $\blacksquare$

C.3 Generalized Frobenius

Proposition C.3. *Under Assumptions 2–5,*

$$\mathcal{L}_n^{\gamma,F}(\Sigma_n,\widetilde{S}_n) \xrightarrow{\text{a.s.}} \int_{-\infty}^{+\infty}\gamma(t)^2\,dH(t) - 2\sum_{k=1}^{\kappa}\int_{a_k}^{b_k}\int_{-\infty}^{+\infty}\gamma(t)\,\gamma\big(\widetilde{\varphi}(x)\big)\cdot\theta(x,t)\,dH(t)\,dF(x) + \sum_{k=1}^{\kappa}\int_{a_k}^{b_k}\gamma\big(\widetilde{\varphi}(x)\big)^2\,dF(x)\ . \tag{C.22}$$

Proof. For simplicity: $\mathsf{Supp}(F)=[a,b]\subset(0,+\infty)$. From Appendix B.13:

$$\begin{aligned}\mathcal{L}_n^{\gamma,\mathrm{F}}(\Sigma_n,\widetilde{S}_n) &= \frac{1}{p}\sum_{i=1}^{p}\{\gamma(\tau_{n,i})^2 - 2u'_{n,i}\gamma(\Sigma_n)u_{n,i}\gamma(\widetilde{d}_{n,i}) + \gamma(\widetilde{d}_{n,i})^2\} \\ &= \frac{1}{p}\sum_{j=1}^{p}\gamma(\tau_{n,j})^2 - \frac{2}{p}\sum_{i=1}^{p}\sum_{j=1}^{p}\gamma(\tau_{n,j})\gamma\big(\widetilde{\varphi}(\lambda_{n,i})\big)\cdot\big(u'_{n,i}v_{n,j}\big)^2 + \frac{1}{p}\sum_{i=1}^{p}\gamma\big(\widetilde{\varphi}(\lambda_{n,i})\big)^2 \\ &= \int_{-\infty}^{+\infty}\gamma(t)^2dH_n(t) - 2\int_a^b\int_{-\infty}^{+\infty}\gamma(t)\gamma\big(\widetilde{\varphi}_n(x)\big)d^2\Theta_n(x,t) + \int_a^b\gamma\big(\widetilde{\varphi}_n(x)\big)^2dF_n(x).\end{aligned} \tag{C.23}$$

By applying the technique from the proof of Theorem 3.1 of Ledoit and Wolf (2018b), and by using Theorem 3.2:

$$\mathcal{L}_n^{\gamma,\mathrm{F}}(\Sigma_n,\widetilde{S}_n) \xrightarrow{\text{a.s.}} \int_{-\infty}^{+\infty}\gamma(t)^2\,dH(t) - 2\sum_{k=1}^{\kappa}\int_{a_k}^{b_k}\int_{-\infty}^{+\infty}\gamma(t)\,\gamma\big(\widetilde{\varphi}(x)\big)\cdot\theta(x,t)\,dH(t)\,dF(x) + \sum_{k=1}^{\kappa}\int_{a_k}^{b_k}\gamma\big(\widetilde{\varphi}(x)\big)^2\,dF(x)\ .\ \blacksquare \tag{C.24}$$

Corollary C.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 (C.22) of the Generalized Frobenius loss if and only if its limiting shrinkage function $\widetilde{\varphi}$ verifies $\forall x\in\mathsf{Supp}(F)$, $\widetilde{\varphi}(x)=\varphi^{\gamma}(x)$, where*

$$\forall x\in\mathsf{Supp}(F)\qquad \varphi^{\gamma}(x) := \gamma^{-1}\left[\int_{-\infty}^{+\infty}\gamma(t)\cdot\theta(x,t)dH(t)\right]\ . \tag{C.25}$$

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

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

$$-2\int_{-\infty}^{+\infty}\gamma(t)\gamma\big(\widetilde{\varphi}(x)\big)\cdot\theta(x,t)\,dH(t) + \gamma\big(\widetilde{\varphi}(x)\big)^2\ . \tag{C.26}$$

The partial derivative of (C.26) with respect to $\widetilde{\varphi}(x)$ is

$$-2\int_{-\infty}^{+\infty}\gamma(t)\gamma'\big(\widetilde{\varphi}(x)\big)\cdot\theta(x,t)\,dH(t) + 2\gamma'\big(\widetilde{\varphi}(x)\big)\gamma\big(\widetilde{\varphi}(x)\big) \tag{C.27}$$

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

$$\widetilde{\varphi}(x) = \gamma^{-1}\left(\int_{-\infty}^{+\infty}\gamma(t)\cdot\theta(x,t)\,dH(t)\right)\ .\ \blacksquare \tag{C.28}$$

The proof of Theorem 4.5 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 $\widehat{H}_n(t) \coloneqq \sum_{i=j}^{p} \mathbb{1}_{[\widehat{\tau}_{n,j},+\infty]}(t)/p$; and to the unobservable quantity $\theta(x)$ corresponds the plug-in estimator $\widehat{\theta}_n(x,t)$ from Equation (3.7). 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 Ledoit and Wolf's (2018b) Theorem 5.2. ■

C.4 Generalized Kullback-Leibler Divergence

Proposition C.4. *Under Assumptions 2–5,*

$$\begin{aligned}\mathcal{L}_n^{\gamma,KL}(\Sigma_n, \widetilde{S}_n) \xrightarrow{\text{a.s.}} & \frac{1}{2}\sum_{k=1}^{\kappa}\int_{a_k}^{b_k}\int_{-\infty}^{+\infty}\frac{\gamma(t)}{\gamma\big(\widetilde{\varphi}(x)\big)}\cdot\theta(x,t)\,dH(t)\,dF(x) \\ & + \frac{1}{2}\sum_{k=1}^{\kappa}\int_{a_k}^{b_k}\log\big[\gamma\big(\widetilde{\varphi}(x)\big)\big]\;dF(x) \\ & - \frac{1}{2}\int_{-\infty}^{+\infty}\log\big[\gamma(t)\big]\,dH(t) - \frac{1}{2}\;. \end{aligned} \tag{C.29}$$

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

$$\mathcal{L}_n^{\gamma,\mathrm{KL}}(\Sigma_n, \widetilde{S}_n) = \frac{1}{2p}\sum_{i=1}^{p}\left\{\frac{u_i'\gamma(\Sigma)u_i}{\gamma\big(\widetilde{\delta}_i\big)} + \log\Big[\gamma\big(\widetilde{\delta}_i\big)\Big] - \log\big[\gamma\big(\tau_i\big)\big] - 1\right\} \tag{C.30}$$

$$\begin{aligned} &= \frac{1}{2}\int_a^b\int_{-\infty}^{+\infty}\frac{\gamma(t)}{\gamma\big(\widetilde{\varphi}_n(x)\big)}\,d^2\Theta_n(x,t) + \frac{1}{2}\int_a^b\log\big[\gamma\big(\widetilde{\varphi}_n(x)\big)\big]\,dF_n(x) \\ &\quad - \frac{1}{2}\int_{-\infty}^{+\infty}\log\big[\gamma(t)\big]dH_n(t) - \frac{1}{2}. \end{aligned} \tag{C.31}$$

By applying the technique from the proof of Theorem 3.1 of Ledoit and Wolf (2018b), and by using Theorem 3.2:

$$\begin{aligned}\mathcal{L}_n^{\gamma,\mathrm{KL}}(\Sigma_n, \widetilde{S}_n) \xrightarrow{\text{a.s.}} & \frac{1}{2}\sum_{k=1}^{\kappa}\int_{a_k}^{b_k}\int_{-\infty}^{+\infty}\frac{\gamma(t)}{\gamma\big(\widetilde{\varphi}(x)\big)}\cdot\theta(x,t)\,dH(t)\,dF(x) \\ & + \frac{1}{2}\sum_{k=1}^{\kappa}\int_{a_k}^{b_k}\log\big[\gamma\big(\widetilde{\varphi}(x)\big)\big]\;dF(x) \\ & - \frac{1}{2}\int_{-\infty}^{+\infty}\log\big[\gamma(t)\big]\,dH(t) - \frac{1}{2}\;. \end{aligned} \tag{C.32}$$

Corollary C.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 (C.29) of the Generalized Kullback-Leibler loss if and only if its limiting shrinkage function $\widetilde{\varphi}$ verifies $\forall x \in \mathsf{Supp}(F)$, $\widetilde{\varphi}(x) = \varphi^{\gamma}(x)$, where $\varphi^{\gamma}(x)$ is defined by Equation (C.25). This results in the same oracle covariance matrix estimator $S_n^{\gamma} \coloneqq U_n\mathsf{Diag}\big(\varphi^{\gamma}(\lambda_{n,1}),\ldots,\varphi^{\gamma}(\lambda_{n,p})\big)U_n'$ as in Corollary C.3.*

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

$$\frac{1}{2}\int_{-\infty}^{+\infty} \frac{\gamma(t)}{\gamma\big(\widetilde{\varphi}(x)\big)} \cdot \theta(x,t)\, dH(t) + \log\big[\gamma\big(\widetilde{\varphi}(x)\big)\big] \ . \tag{C.33}$$

The partial derivative of (C.33) with respect to $\widetilde{\varphi}(x)$ is

$$-\frac{1}{2}\int_{-\infty}^{+\infty} \frac{\gamma(t)}{\gamma\big(\widetilde{\varphi}(x)\big)^2}\gamma'\big(\widetilde{\varphi}(x)\big) \cdot \theta(x,t)\, dH(t) + \frac{\gamma'\big(\widetilde{\varphi}(x)\big)}{2\gamma\big(\widetilde{\varphi}(x)\big)} \tag{C.34}$$

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

$$\widetilde{\varphi}(x) = \gamma^{-1}\left(\int_{-\infty}^{+\infty} \gamma(t) \cdot \theta(x,t)\, dH(t)\right) \ . \blacksquare \tag{C.35}$$

The proof of Theorem 4.6 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. ■

C.5 Singular Case

There is one small difference with the non-singular case: the support of the limiting sample spectral distribution F is now $\mathsf{Supp}(F) = \{0\} \cup (\bigcup_{k=1}^{\kappa}[a_k, b_k])$, where (as before) $0 < a_1 < b_1 < \cdots < a_\kappa < b_\kappa < \infty$.

Proposition C.5. *Under Assumptions 3–6,*

$$\begin{aligned}\mathcal{L}_n^{\gamma,F}(\Sigma_n, \widetilde{S}_n) \xrightarrow{\text{a.s.}} &\int_{-\infty}^{+\infty} \gamma(t)^2\, dH(t) - 2\sum_{k=1}^{\kappa}\int_{a_k}^{b_k}\int_{-\infty}^{+\infty} \gamma(t)\,\gamma\big(\widetilde{\varphi}(x)\big) \cdot \theta(x,t)\, dH(t)\, dF(x) \\ &- 2\frac{c-1}{c}\int_{-\infty}^{+\infty} \gamma(t)\,\gamma\big(\widetilde{\varphi}(0)\big) \cdot \theta(0,t)\, dH(t) \\ &+ \sum_{k=1}^{\kappa}\int_{a_k}^{b_k} \gamma\big(\widetilde{\varphi}(x)\big)^2\, dF(x) + \frac{c-1}{c}\gamma\big(\widetilde{\varphi}(0)\big)^2 \ ,\end{aligned} \tag{C.36}$$

where $\theta(0,t)$ is given by Equation (5.4).

Proof. For simplicity, we assume that $\mathsf{Supp}(F) = \{0\} \cup [a,b]$, as the extension to the case $\kappa > 1$

is straightforward. Starting from the proof of Proposition C.3:

$$\begin{aligned}
\mathcal{L}_n^{\gamma,\mathrm{F}}(\Sigma_n, \widetilde{S}_n) &= \frac{1}{p}\sum_{j=1}^{p}\gamma(\tau_{n,j})^2 - \frac{2}{p}\sum_{i=1}^{p}\sum_{j=1}^{p}\gamma(\tau_{n,j})\gamma\big(\widetilde{\varphi}_n(\lambda_{n,i})\big)\cdot\big(u_{n,i}'v_{n,j}\big)^2 + \frac{1}{p}\sum_{i=1}^{p}\gamma\big(\widetilde{\varphi}_n(\lambda_{n,i})\big)^2 \\
&= \frac{1}{p}\sum_{j=1}^{p}\gamma(\tau_{n,j})^2 - \frac{2}{p}\sum_{i=p-n+1}^{p}\sum_{j=1}^{p}\gamma(\tau_{n,j})\gamma\big(\widetilde{\varphi}_n(\lambda_{n,i})\big)\cdot\big(u_{n,i}'v_{n,j}\big)^2 \\
&\quad - \frac{2}{p}\sum_{i=1}^{p-n}\sum_{j=1}^{p}\gamma(\tau_{n,j})\gamma\big(\widetilde{\varphi}_n(0)\big)\cdot\big(u_{n,i}'v_{n,j}\big)^2 \\
&\quad + \frac{1}{p}\sum_{i=p-n+1}^{p}\gamma\big(\widetilde{\varphi}_n(\lambda_{n,i})\big)^2 + \frac{1}{p}\sum_{i=1}^{p-n}\gamma\big(\widetilde{\varphi}_n(0)\big)^2 \qquad \text{(C.37)} \\
&= \int_{-\infty}^{+\infty}\gamma(t)^2 dH_n(t) - 2\int_a^b\int_{-\infty}^{+\infty}\gamma(t)\gamma\big(\widetilde{\varphi}_n(x)\big)d^2\Theta_n(x,t) \\
&\quad - 2\frac{p-n}{p}\int_{-\infty}^{+\infty}\gamma(t)\gamma\big(\widetilde{\varphi}_n(0)\big)d\Theta_n(0,t) \\
&\quad + \int_a^b\gamma\big(\widetilde{\varphi}_n(x)\big)^2 dF_n(x) + \frac{p}{p-n}\gamma\big(\widetilde{\varphi}_n(0)\big)^2\,. \qquad \text{(C.38)}
\end{aligned}$$

By applying the technique from the proof of Theorem 6.1 of Ledoit and Wolf (2018b), and by using Theorem 3 of Ledoit and Péché (2011) to handle the limit of Θ_n, it follows that:

$$\begin{aligned}
\mathcal{L}_n^{\gamma,\mathrm{F}}(\Sigma_n, \widetilde{S}_n) \xrightarrow{\text{a.s.}} &\int_{-\infty}^{+\infty}\gamma(t)^2\,dH(t) - 2\sum_{k=1}^{\kappa}\int_{a_k}^{b_k}\int_{-\infty}^{+\infty}\gamma(t)\,\gamma\big(\widetilde{\varphi}(x)\big)\cdot\theta(x,t)\,dH(t)\,dF(x) \\
&- 2\frac{c-1}{c}\int_{-\infty}^{+\infty}\gamma(t)\,\gamma\big(\widetilde{\varphi}(0)\big)\cdot\frac{1}{\big(1-\frac{1}{c}\big)\big[1+\breve{m}_{\underline{F}}(0)\,t\big]}\,dH(t) \\
&+ \sum_{k=1}^{\kappa}\int_{a_k}^{b_k}\gamma\big(\widetilde{\varphi}(x)\big)^2\,dF(x) + \frac{c-1}{c}\gamma\big(\widetilde{\varphi}(0)\big)^2\,. \blacksquare \qquad \text{(C.39)}
\end{aligned}$$

Corollary C.5. *Under Assumptions 3–6, a covariance matrix estimator $\widetilde{S}_n$ minimizes in the class of rotation-equivariant estimators the a.s. limit (C.36) of the Generalized Frobenius loss if and only if its limiting shrinkage function $\widetilde{\varphi}$ verifies $\forall x \in \mathsf{Supp}(F)$, $\widetilde{\varphi}(x) = \varphi^{\gamma}(x)$, where*

$$\forall x \in \mathsf{Supp}(F) \qquad \varphi^{\gamma}(x) := \gamma^{-1}\left[\int_{-\infty}^{+\infty}\gamma(t)\cdot\theta(x,t)dH(t)\right]\,. \qquad \text{(C.40)}$$

This yields the oracle covariance matrix estimator $S_n^{\gamma} := U_n\mathsf{Diag}\big(\varphi^{\gamma}(\lambda_{n,1}),\ldots,\varphi^{\gamma}(\lambda_{n,p})\big)U_n'$.

For $x \in \bigcup_{k=1}^{\kappa}[a_k, b_k]$, the proof is the same as for Corollary C.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

$$-2\frac{c-1}{c}\int_{-\infty}^{+\infty}\gamma(t)\gamma\big(\widetilde{\varphi}(0)\big)\cdot\theta(0,t)\,dH(t) + \frac{c-1}{c}\gamma\big(\widetilde{\varphi}(0)\big)^2\,. \qquad \text{(C.41)}$$

The partial derivative of (C.41) with respect to $\widetilde{\varphi}(0)$ is

$$-2\frac{c-1}{c}\int_{-\infty}^{+\infty}\gamma(t)\gamma'\big(\widetilde{\varphi}(0)\big)\cdot\theta(0,t)\,dH(t) + 2\frac{c-1}{c}\gamma'\big(\widetilde{\varphi}(0)\big)\gamma\big(\widetilde{\varphi}(0)\big) \qquad \text{(C.42)}$$

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

$$\widetilde{\varphi}(0) = \gamma^{-1}\left(\int_{-\infty}^{+\infty} \gamma(t) \cdot \theta(0,t)\, dH(t)\right) . \blacksquare \tag{C.43}$$

Proof of Theorem 5.1. The proof 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 $\widehat{H}_n(t) \coloneqq \sum_{i=j}^{p} \mathbb{1}_{[\widehat{\tau}_{n,j},+\infty]}(t)/p$; to the unobservable function $\theta(x,t)$ corresponds the plug-in estimator $\widehat{\theta}_n(x,t)$ from Equation (3.7) for $x > 0$; and to the unobservable quantity $\theta(0,t)$ corresponds the plug-in estimator $\widehat{\theta}_n(0,t)$ from Equation (5.5). 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 Ledoit and Wolf's (2018b) Theorem 6.2. ■