The Annals of Statistics
2020, Vol. 48, No. 5, 3043–3065
https://doi.org/10.1214/19-AOS1921

ANALYTICAL NONLINEAR SHRINKAGE OF LARGE-DIMENSIONAL COVARIANCE MATRICES

BY OLIVIER LEDOIT[*] AND MICHAEL WOLF[**]

Department of Economics, University of Zurich, [*]olivier.ledoit@econ.uzh.ch; [**]michael.wolf@econ.uzh.ch

This paper establishes the first analytical formula for nonlinear shrinkage estimation of large-dimensional covariance matrices. We achieve this by identifying and mathematically exploiting a deep connection between nonlinear shrinkage and nonparametric estimation of the Hilbert transform of the sample spectral density. Previous nonlinear shrinkage methods were of numerical nature: QuEST requires numerical inversion of a complex equation from random matrix theory whereas NERCOME is based on a sample-splitting scheme. The new analytical method is more elegant and also has more potential to accommodate future variations or extensions. Immediate benefits are (i) that it is typically 1000 times faster with basically the same accuracy as QuEST and (ii) that it accommodates covariance matrices of dimension up to 10,000 and more. The difficult case where the matrix dimension exceeds the sample size is also covered.

1. Introduction.

Given that many researchers employ the linear shrinkage estimator of Ledoit and Wolf (2004) to estimate covariance matrices whose dimensions, p, are commensurate with the sample size, n, attention is naturally turning to the more difficult—but potentially more rewarding—approach of *nonlinear* shrinkage estimation, where the transformation applied to the sample eigenvalues must be optimal not in a space of dimension two (intercept and slope) but in a much larger space of dimension p (i.e., unconstrained transformation).

So far, there exist two very different nonlinear shrinkage methods that give satisfactory and largely compatible results. The first method is the *indirect* approach of Ledoit and Wolf (2012, 2015). It is indirect because it goes through recovery of the population eigenvalues. They are not a necessary part of the procedure and are notoriously hard to pin down, so they can be thought of as *nuisance* parameters. The method relies on numerical inversion of a deterministic multivariate function called the Quantized Eigenvalues Sampling Transform (QuEST) function, which essentially maps population eigenvalues into sample eigenvalues. The mathematics come from the field known as random matrix theory, originally from physics, and involve heavy usage of integral transforms.

The second method, going back to Abadir, Distaso and Žikeš (2014), is much simpler conceptually. It involves just splitting the sample into two parts: one to estimate the eigenvectors, and the other to estimate the eigenvalues associated with these eigenvectors. Averaging over a large number of permutations of the sample split makes the method perform well. Lam (2016) calls this method Nonparametric Eigenvalue-Regularized COvariance Matrix Estimator (NERCOME). In practice, it requires brute-force spectral decomposition of many different large-dimensional matrices. The main attraction of NERCOME lies not in the fact that it would be more accurate or faster, but in the fact that it is decisively simpler and more transparent, thus providing an independent and easily verifiable confirmation for the mathematically delicate indirect method of QuEST.

Received November 2018; revised October 2019.
MSC2020 subject classifications. Primary 62H12; secondary 62G20, 15A52.
Key words and phrases. Hilbert transform, kernel estimation, rotation equivariance.

The goal of this paper is to develop a method that combines the best qualities of the three approaches described above: the speed of linear shrinkage, the accuracy of the QuEST function and the transparency of NERCOME. We achieve this goal through nonparametric kernel estimation of the limiting spectral density of the sample eigenvalues *and* its Hilbert transform. From the QuEST route, we borrow the optimal nonlinear shrinkage formula; from NERCOME, we imitate the simplicity of interpretation and code (we need just over 20 lines in Matlab); and from linear shrinkage we borrow the speed, scalability and analytical nature.

We contribute to the existing literature on three levels. At the conceptual level, we show how the presence of the Hilbert transform in the shrinkage formula is the ingredient that induces "shrinkage" by attracting nearby eigenvalues toward each other, thereby reducing cross-sectional dispersion. The Hilbert transform is also what makes shrinkage a local (as opposed to global) phenomenon, which explains why there are nonlinearities. At the technical level, we extend the kernel estimator of the limiting spectral density function of large-dimensional sample covariance matrices developed by Jing et al. (2010) in two important directions. First, we estimate not just the density but also its Hilbert transform; indeed, from the point of view of optimal covariance matrix estimation, the Hilbert transform is equally as important as the density itself. Krantz ((2009), page 17) alludes to this importance being commonplace in mathematics: "The Hilbert transform is, without question, the most important operator in analysis. It arises in many different contexts, and all these contexts are intertwined in profound and influential ways." Our second extension of the kernel estimator is that, instead of keeping the bandwidth constant (or uniform) for a given sample size, we let it vary in proportion to the location of a given sample eigenvalue. This improvement confines the support of the spectral density estimator to the positive half of the real line, as befits positive-definite matrices; it also reflects the scale-invariance of the problem. Finally, at the operational level, we make the computer code two orders of magnitude simpler and faster than the indirect route of numerically inverting the QuEST function. As a result, we can estimate covariance matrices of dimension 10,000 and beyond, whereas the largest magnitude that could be handled by nonlinear shrinkage before was 1000.

The remainder of the paper is organized as follows. Section 2 describes within a finite-sample framework the basic features of the estimation problem under consideration. Section 3 moves it to the realm of large-dimensional asymptotics and establishes necessary background. Section 4 develops our proportional-bandwidth estimator for the limiting sample spectral density and its Hilbert transform. Section 5 runs an extensive set of Monte Carlo simulations. Section 6 concludes. The Supplementary Material (Ledoit and Wolf (2020)) contains all mathematical proofs, further simulation results, the extension to the singular case, various robustness checks and our code.

2. Finite samples. In this section, and this section only, the sample size, n, and the covariance matrix dimension, p, are fixed for expositional purposes. Even though n is temporarily fixed, we still subscript the major objects with n in order to maintain compatibility of notation with the subsequent sections that let n (and p) go to infinity under large-dimensional asymptotics.

2.1. *Rotation equivariance.* Let Σ_n denote a p-dimensional population covariance matrix. A mean-zero independent and identically distributed (i.i.d.) sample of n observations Y_n generates the sample covariance matrix $S_n := Y_n' Y_n / n$. Its spectral decomposition is $S_n = U_n \Lambda_n U_n'$, where Λ_n is the diagonal matrix, whose elements are the eigenvalues $\boldsymbol{\lambda}_n := (\lambda_{n,1}, \dots, \lambda_{n,p})$ sorted in nondecreasing order without loss of generality, and an orthogonal matrix U_n whose columns $[u_{n,1} \cdots u_{n,p}]$ are the corresponding eigenvectors. We seek an estimator of the form $\widehat{\Sigma}_n := U_n \widehat{\Delta}_n U_n'$, where $\widehat{\Delta}_n$ is a diagonal matrix whose elements $\widehat{\boldsymbol{\delta}}_n := (\widehat{\delta}_{n,1}, \dots, \widehat{\delta}_{n,p}) \in (0, +\infty)^p$ are a function of $\boldsymbol{\lambda}_n$. Thus, $\widehat{\Sigma}_n = \sum_{i=1}^p \widehat{\delta}_{n,i} \cdot u_{n,i} u_{n,i}'$.

This is the framework of rotation equivariance championed by Stein ((1986), Lecture 4). Rotating the original set of p variables is viewed as an uninformative linear transformation that must not contaminate the estimation procedure. The underlying philosophy is that all orthonormal bases of the Euclidian space $\mathbb{R}^p$ are equivalent. By contrast, in the sparsity literature, the original basis is special because a matrix that is sparse in the original basis is generally no longer sparse in any other basis. Rotation equivariance does not take a stance on the orientation of the eigenvectors of the population covariance matrix.

REMARK 2.1. To simplify the notation, we assume that all variables have mean zero. In many applications, variables do not have mean zero, or at least it is not known whether they do. In such a setting, it is more common to base the sample covariance matrix on the demeaned data instead: $S_n := \widetilde{Y}_n' \widetilde{Y}_n/(n-1)$, where $\widetilde{Y}_n$ is obtained from Y_n by the operation of columnwise demeaning. In this case, n needs to be replaced everywhere with the "effective" sample size $n-1$. As shown at the beginning of Section 3 of Silverstein and Bai (1995), demeaning is a rank-one perturbation which in turn, thanks to Lemma 2.5a of the same paper, implies that it has no impact on large-dimensional asymptotic convergence results.

2.2. *Loss function.* A perennial question is how to quantify the usefulness of a covariance matrix estimator. It devolves into asking what covariance matrix estimators are used for. They are often used to find combinations of the original variables that have *minimum variance* under a linear constraint. Important—and mathematically equivalent—examples include Markowitz (1952) portfolio selection in finance, Capon (1969) beamforming in signal processing and optimal fingerprinting (Ribes, Azaïs and Planton (2009)) in climate research. The quality of the covariance matrix estimator is then measured by the *true* variance of the linear combination of the original variables: lower variance is better.

On this basis, a metric that is agnostic as to the actual orientation of the linear constraint vector, and is justified under large-dimensional asymptotics, has been proposed by Engle, Ledoit and Wolf ((2019), Definition 1). It can be expressed in our notation as

$$\mathcal{L}_n^{\mathrm{MV}}(\widehat{\Sigma}_n, \Sigma_n) := \frac{\mathsf{Tr}(\widehat{\Sigma}_n^{-1} \Sigma_n \widehat{\Sigma}_n^{-1})/p}{[\mathsf{Tr}(\widehat{\Sigma}_n^{-1})/p]^2} - \frac{1}{\mathsf{Tr}(\Sigma_n^{-1})/p}, \tag{2.1}$$

where $\mathsf{Tr}(\cdot)$ denotes the trace of a square matrix. $\mathcal{L}_n^{\mathrm{MV}}$ represents the *true* variance of the linear combination of the original variables that has the minimum *estimated* variance, under a generic linear constraint, after suitable normalization. Further justification for the minimum variance (MV) loss function is provided by Engle and Colacito (2006) and Ledoit and Wolf (2017a). The optimal nonlinear shrinkage formula in finite samples is identified by the following proposition.

PROPOSITION 2.1. *An estimator* $\widehat{\Sigma}_n := \sum_{i=1}^p \widehat{\delta}_{n,i} \cdot u_{n,i} u_{n,i}'$ *minimizes the MV loss function* $\mathcal{L}_n^{\mathrm{MV}}$ *defined in equation* (2.1) *within the class of rotation-equivariant estimators specified in Section* 2.1 *if and only if there exists a scalar* $\beta_n \in (0, +\infty)$ *such that* $\widehat{\delta}_{n,i} = \beta_n \cdot u_{n,i}' \Sigma_n u_{n,i}$, *for* $i = 1, \dots, p$.

Among all the possible scaling factors $\beta_n \in (0, +\infty)$, the default value $\beta_n = 1$ will be retained from here onward because $\sum_{i=1}^p u_{n,i}' \Sigma_n u_{n,i} = \mathsf{Tr}(\Sigma_n)$. Thus, finite-sample optimal nonlinear shrinkage replaces the sample eigenvalues $\boldsymbol{\lambda}_n$ with the unobservable quantity

$$\boldsymbol{d}_n^* := (d_{n,1}^*, \dots, d_{n,p}^*) := (u_{n,1}' \Sigma_n u_{n,1}, \dots, u_{n,p}' \Sigma_n u_{n,p}), \tag{2.2}$$

prior to recombining it with the sample eigenvectors to form the (nonfeasible) covariance matrix estimator

$$S_n^* := \sum_{i=1}^{p} d_{n,i}^* \cdot u_{n,i} u_{n,i}' = \sum_{i=1}^{p} \bigl(u_{n,i}' \Sigma_n u_{n,i}\bigr) \cdot u_{n,i} u_{n,i}'. \tag{2.3}$$

REMARK 2.2. Section 3.1 of Ledoit and Wolf (2012) shows that the same estimator S_n^* is also finite-sample optimal with respect to the (squared) Frobenius loss function, which is defined for a generic estimator $\widehat{\Sigma}_n$ as

$$\mathcal{L}_n^{\mathrm{FR}}(\widehat{\Sigma}_n, \Sigma_n) := \frac{1}{p} \mathsf{Tr}\bigl[(\widehat{\Sigma}_n - \Sigma_n)^2\bigr]. \tag{2.4}$$

This is the loss function with respect to which Ledoit and Wolf's (2004) linear shrinkage estimator is optimized. Appendix B in the Supplementary Material (Ledoit and Wolf (2020)) contains corresponding Monte Carlo simulations.

3. Large-dimensional asymptotics. Further investigations of the nonlinear shrinkage formula that maps $\boldsymbol{\lambda}_n$ into $\boldsymbol{d}_n^*$ are mathematically arduous or even unattainable in finite samples, but decisive progress can be made by letting the dimension go to infinity together with the sample size.

3.1. *Assumptions.* The major assumptions that define the large-dimensional asymptotic framework are listed below. They are similar, for example, to the ones made by Ledoit and Wolf (2018).

ASSUMPTION 3.1 (Dimension). Let n denote the sample size and $p := p(n)$ the number of variables. It is assumed that the "concentration (ratio)" $c_n := p/n$ converges, as $n \to \infty$, 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.

The case $c > 1$, where the sample covariance matrix is singular, is covered in Appendix C in the Supplementary Material (Ledoit and Wolf (2020)). The case $c = 1$ is not covered by the mathematical (random matrix) theory but is addressed via Monte Carlo simulations at the end of Appendix C as well.

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

ASSUMPTION 3.2 (Population covariance matrix).

a. The population covariance matrix Σ_n is a nonrandom symmetric positive-definite matrix of dimension $p \times p$.

b. Let $\boldsymbol{\tau}_n := (\tau_{n,1}, \dots, \tau_{n,p})'$ denote a system of eigenvalues of Σ_n, and H_n the e.d.f. of population eigenvalues. It is assumed that H_n converges weakly to a limit law H, called the "limiting spectral distribution (function)."

c. $\mathsf{Supp}(H)$, the support of H, is the union of a finite number of closed intervals, bounded away from zero and infinity.

d. There exists a compact interval $[\underline{T}, \overline{T}] \subset (0, \infty)$ that contains $\{\tau_{n,1}, \dots, \tau_{n,p}\}$ for all n large enough.

ASSUMPTION 3.3 (Data generating process). X_n is a $n \times p$ matrix of i.i.d. random variables with mean zero, variance one and finite 16th moment. The matrix of observations is $Y_n := X_n \times \sqrt{\Sigma_n}$. Neither $\sqrt{\Sigma_n}$ nor X_n are observed on their own: only Y_n is observed.

REMARK 3.1. The assumption of finite 16th moment is used in Theorem 3 of Jing et al. (2010), which we will utilize in the proof of our own Theorem 4.1. However, these authors' Remark 1 conjectures that a finite 4th moment is enough, which is supported by Monte Carlo simulations we report in Table 4.

The sample covariance matrix S_n, its eigenvalues $\boldsymbol{\lambda}_n := (\lambda_{n,1}, \dots, \lambda_{n,p})$ and eigenvectors $U_n := [u_{n,1} \cdots u_{n,p}]$ have already been defined in Section 2.1. The e.d.f. of sample eigenvalues is the function $F_n(x) := \sum_{i=1}^{p} \mathbb{1}_{\{\lambda_{n,i} \le x\}}/p$ for $x \in \mathbb{R}$.

3.2. *Random matrix theory.* The literature on the limiting behavior of the eigenvalues of the sample covariance matrix under large-dimensional asymptotics is based on a foundational result by Marčenko and Pastur (1967). It has been strengthened and broadened by subsequent authors including Silverstein and Bai (1995) and Silverstein (1995), among others. The latter's Theorem 1.1 implies that, under Assumptions 3.1–3.3, there exists a limiting sample spectral distribution F such that $\forall x \in \mathbb{R}$, $F_n(x) \xrightarrow{\text{a.s.}} F(x)$. This limiting distribution F is uniquely determined by c and H; therefore, we will refer to it as $F_{c,H} := F$ whenever clarification is needed.

Assumptions 3.1–3.3 together with Theorem 1.1 of Bai and Silverstein (1998) imply that the support of F, denoted by $\mathsf{Supp}(F)$, is the union of a finite number $\nu \ge 1$ of compact intervals: $\mathsf{Supp}(F) = \bigcup_{k=1}^{\nu} [a_k, b_k]$, where $0 < a_1 < b_1 < \cdots < a_\nu < b_\nu < \infty$.

3.3. *Hilbert transform.* At this point, it is necessary to introduce an important mathematical tool called the *Hilbert transform*, which is defined as convolution with the *Cauchy kernel* $\frac{dt}{\pi t}$.

DEFINITION 3.2. The Hilbert transform of a real function g is defined as

$$\forall x \in \mathbb{R} \quad \mathcal{H}_g(x) := \frac{1}{\pi} \operatorname{PV} \int_{-\infty}^{+\infty} g(t) \frac{dt}{t - x}. \tag{3.1}$$

Here, PV denotes the *Cauchy principal value*, which is used to evaluate the singular integral in the following way:

$$\operatorname{PV} \int_{-\infty}^{+\infty} g(t) \frac{dt}{t - x} := \lim_{\varepsilon \to 0^+} \left[\int_{-\infty}^{x-\varepsilon} g(t) \frac{dt}{t - x} + \int_{x+\varepsilon}^{+\infty} g(t) \frac{dt}{t - x} \right]. \tag{3.2}$$

Recourse to the Cauchy principal value is needed because the Cauchy kernel is singular, as a consequence of which the integral does not converge in the usual sense.

The intuition behind the Hilbert transform is that it operates like a local attraction force. It is very positive if there are heavy mass points slightly larger than you, so it pushes you up (toward them), but very negative if they are slightly smaller, so it pushes you down (*also* toward them). When the mass points lie far away, it fades out to zero like gravitational attraction does. These effects can be deduced simply by visual inspection of the Cauchy kernel. Figure 1 confirms them visually by plotting the Hilbert transform of four well-known densities.

Obviously, the regularity of the Hilbert transform is a direct reflection of the regularity of the underlying density, but the main effects as described above remain true across the board. The formulas used in Figure 1 come from Erdélyi et al. ((1954), Chapter XV); for convenience, they are reproduced in Table 1.

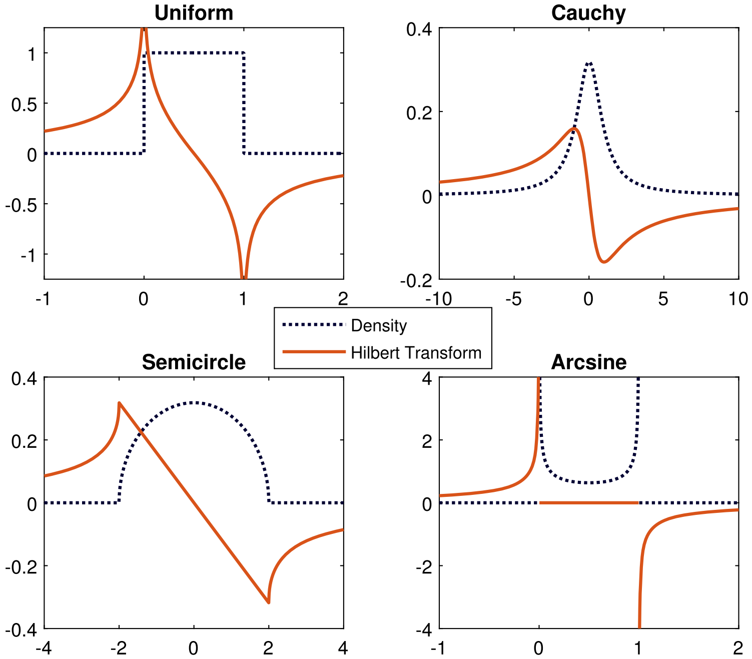

FIG. 1. *Hilbert transform of four densities. The transform is strongly positive to the left of the center of mass, strongly negative to the right and vanishes away from the center of mass.*

Theorem 1.1 of Silverstein and Choi (1995) shows that the limiting spectral density $f := F'$ exists and is continuous, and that its Hilbert transform $\mathcal{H}_f$ exists and is continuous, too. As we shall see below, f and $\mathcal{H}_f$ are the two key ingredients in computing the optimal nonlinear shrinkage formula.

REMARK 3.2. The reason why we use the Hilbert transform is because it is the real part of the extension to the real line of the Stieltjes (1894) transform, and the vast majority of the known results on large-dimensional sample covariance matrix eigenvalues have been couched in terms of the Stieltjes transform ever since the seminal paper of Marčenko and Pastur (1967). We could have written the whole paper in terms of the Stieltjes transform instead of the Hilbert transform, but we figured that avoiding an excursion into the complex plane was clearer and more economical.

TABLE 1
Formulas for various densities and their Hilbert transforms

	Density	Hilbert transform
Uniform	$f(x)=\mathbb{1}_{\{0\le x<1\}}$	$\mathcal{H}_f(x)=\frac{1}{\pi}\log\lvert\frac{1-x}{x}\rvert$
Cauchy	$f(x)=\frac{1}{\pi(x^2+1)}$	$\mathcal{H}_f(x)=-\frac{x}{\pi(x^2+1)}$
Semicircle	$f(x)=\frac{\sqrt{\max\{4-x^2,0\}}}{2\pi}$	$\mathcal{H}_f(x)=\frac{-x+\operatorname{sgn}(x)\sqrt{\max\{x^2-4,0\}}}{2\pi}$
Arcsine	$f(x)=\begin{cases}0, & x<0,\\ \frac{1}{\pi\sqrt{x(1-x)}}, & x\in(0,1),\\ 0, & x>1\end{cases}$	$H_f(x)=\begin{cases}\frac{1}{\pi\sqrt{x(x-1)}}, & x<0,\\ 0, & x\in(0,1),\\ -\frac{1}{\pi\sqrt{x(x-1)}}, & x>1\end{cases}$

3.4. *Optimal nonlinear shrinkage formula.* We consider the same class of nonlinear shrinkage estimators as Ledoit and Wolf (2017a). It constitutes the large-dimensional asymptotic counterpart to the class of rotation-equivariant covariance matrix estimators introduced in Section 2.1.

DEFINITION 3.3 (Class of estimators). Covariance matrix estimators are of the type $\widehat{\Sigma}_n := U_n \widehat{\Delta}_n U_n'$, where $\widehat{\Delta}_n$ is a diagonal matrix: $\widehat{\Delta}_n := \mathsf{Diag}(\widehat{\delta}_n(\lambda_{n,1}), \ldots, \widehat{\delta}_n(\lambda_{n,p}))$, and $\widehat{\delta}_n$ is a (possibly random) real univariate function which can depend on S_n.

The shrinkage function must be as well behaved asymptotically as the population spectral e.d.f.

ASSUMPTION 3.4 (Limiting shrinkage function). There exists a nonrandom real univariate function $\widehat{\delta}$ defined on $\mathsf{Supp}(F)$ and continuously differentiable such that $\widehat{\delta}_n(x) \xrightarrow{\text{a.s}} \widehat{\delta}(x)$, for all $x \in \mathsf{Supp}(F)$. Furthermore, this convergence is uniform over $x \in \bigcup_{k=1}^{\nu}[a_k + \eta, b_k - \eta]$, for any small $\eta > 0$. Finally, for any small $\eta > 0$, there exists a finite nonrandom constant $\widehat{K}$ such that almost surely, over the set $x \in \bigcup_{k=1}^{\nu}[a_k - \eta, b_k + \eta]$, $\widehat{\delta}_n(x)$ is uniformly bounded by $\widehat{K}$ from above and by $1/\widehat{K}$ from below, for all n large enough.

Within this framework, the asymptotically optimal nonlinear shrinkage formula is known.

THEOREM 3.1. *Define the oracle nonlinear shrinkage function*

$$\forall x \in \mathsf{Supp}(F) \quad d^{\mathrm{o}}(x) := \frac{x}{[\pi c x f(x)]^2 + [1 - c - \pi c x \mathcal{H}_f(x)]^2}. \tag{3.3}$$

If Assumptions 3.1–3.4 are satisfied, then the following statements hold true:

(a) *The oracle estimator of the covariance matrix*

$$S_n^{\mathrm{o}} := U_n D_n^{\mathrm{o}} U_n' \quad \textit{where } D_n^{\mathrm{o}} := \mathsf{Diag}\big(d^{\mathrm{o}}(\lambda_{n,1}), \ldots, d^{\mathrm{o}}(\lambda_{n,p})\big) \tag{3.4}$$

minimizes in the class of nonlinear shrinkage estimators defined in Assumption 3.4 the almost sure limit of the minimum variance loss function introduced in Section 2.2, as p and n go to infinity together in the manner of Assumption 3.1.

(b) *Conversely, any covariance matrix estimator $\widehat{\Sigma}_n$ that minimize the a.s. limit of the minimum-variance loss function (2.1) is asymptotically equivalent to S_n^{o} up to scaling, in the sense that its limiting shrinkage function is of the form $\widehat{\delta} = \alpha\, d^{\mathrm{o}}$ for some positive constant α.*

The scaling factor α in part (b) will henceforth be set equal to one in order to ensure that the estimator has the same trace as the sample covariance matrix and the population covariance matrix asymptotically.

Note that S_n^{o} already represents considerable progress with respect to the finite-sample optimal estimator S_n^* of equation (2.3): We no longer need to know the full population covariance matrix Σ_n (estimating $p(p+1)/2$ parameters is hopeless when p is of the same order of magnitude as n); instead, we just need to know its eigenvalues $\boldsymbol{\tau}_n$ (p parameters only, which is *a priori* extractable from a dataset of dimension $p \times n$). The value of equation (3.3) is that it transforms what was apparently an infeasible problem into one that may become feasible provided proper techniques are deployed, thereby avoiding the “curse of dimensionality.”

REMARK 3.3. The quantities $(d^{\mathrm{o}}(\lambda_{n,1}),\dots,d^{\mathrm{o}}(\lambda_{n,p}))$ represent large-dimensional asymptotic counterparts to the finite-sample optimal quantities $(u'_{n,1}\Sigma_n u_{n,1},\dots,u'_{n,p}\Sigma_n u_{n,p})$ of equation (2.2). Equation (3.3) was first discovered by Ledoit and Péché ((2011), Theorem 3), based on a generalization of the fundamental equation of random matrix theory originally due to Marčenko and Pastur (1967). The formula here is the first one expressed without any reference to complex numbers; previous (mathematically equivalent) versions used the complex-valued Stieltjes transform instead of the Hilbert transform.

3.5. *Shrinkage as local attraction via the Hilbert transform.* Equation (3.3) may look initially daunting, yet intuition can be gleaned by considering a slight modification of the limiting sample spectral density: $\varphi(x) := \pi x f(x)$. Multiplication by x captures the fact that larger eigenvalues exert more pull than smaller ones, everything else being equal. Qualitatively speaking, φ acts as surrogate for the density f, in the sense that it measures where the influential eigenvalues lie. Its Hilbert transform is $\mathcal{H}_\varphi(x) = 1 + \pi x \mathcal{H}_f(x)$. In terms of the reweighted density function φ, formula (3.3) becomes

$$\forall x \in \mathsf{Supp}(F) \quad d^{\mathrm{o}}(x) = \frac{x}{1 + c^2[\varphi(x)^2 + H_\varphi(x)^2] - 2cH_\varphi(x)}, \tag{3.5}$$

which is easier to interpret. If the limiting concentration ratio c is negligible, then the denominator is close to one, which would mean no shrinkage: This is why the sample covariance matrix works well under traditional (fixed-dimensional) asymptotics. As c increases, however, (noticeable) shrinkage must occur. Let us set aside the term $c^2[\varphi(x)^2 + H_\varphi(x)^2]$ because it is negligible for small c and generally innocuous: given that it is always positive, it only serves to augment the first term 1. The key factor here is sign of the last term $2cH_\varphi(x)$. It works as a local attraction force. From the point of view of any given eigenvalue $\lambda_{n,i}$, if there is a heavy mass of other eigenvalues hovering slightly above, $2cH_\varphi(\lambda_{n,i})$ will be strongly positive, which will push $\lambda_{n,i}$ higher in the direction of its closest and most numerous neighbors. Conversely, if there are many eigenvalues hovering slightly below $\lambda_{n,i}$, then $2cH_\varphi(\lambda_{n,i})$ will be strongly negative, which will pull $\lambda_{n,i}$ lower—also in the direction of its most immediate neighbors. This attraction phenomenon is intrinsically local because the absolute magnitude of the Hilbert transform $H_\varphi(\lambda_{n,i})$ fades away as the other eigenvalues become more distant from $\lambda_{n,i}$.

The local attraction field generated by the Hilbert transform is why we speak of "shrinkage": the spread of covariance matrix eigenvalues reduces when they get closer to one another. Linear shrinkage is handling this effect at the global level, that is, by shrinking all sample eigenvalues toward their grand mean. However, given that we now know that the attraction is essentially a local phenomenon that fades away at great distances, we must shrink any given eigenvalue toward those of its neighbors that exert the greatest pull. Thus, it could be that it is optimal to nonlinearly "shrink" a relatively small eigenvalue (i.e., one that is below average) downward, if there is a sufficiently massive cluster of slightly inferior eigenvalues attracting it toward them, which cannot happen with linear shrinkage instead: Any eigenvalue below average will be brought up necessarily. Figure 2 provides a graphical illustration of these contrasting behaviors.

In this example, the average eigenvalue is equal to 1.25. Sample eigenvalues below the average but above 1 need to be "shrunk" downward because they are attracted by the cluster to their immediate left. Similarly, sample eigenvalues above the average but below 1.75 need to be "shrunk" upwards because they are attracted by the cluster to their immediate right. Linear shrinkage, being a global operator, is not equipped to sense a disturbance in the force: It applies the same shrinkage intensity across the board and shrinks all eigenvalues toward the average of 1.25.

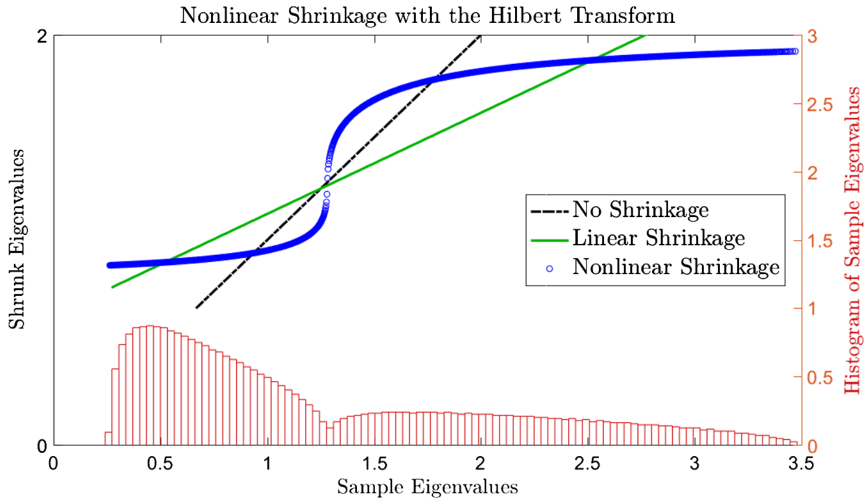

FIG. 2. *Local attraction effect.* 2500 *population eigenvalues are equal to* 0.8, *and* 1500 *are equal to* 2. *The sample size is* $n = 18{,}000$. *At the bottom of the figure is a histogram displaying the location of the sample eigenvalues.*

3.6. *Practical considerations.* $\boldsymbol{d}_n^{\mathrm{o}} := (d^{\mathrm{o}}(\lambda_{n,1}), \dots, d^{\mathrm{o}}(\lambda_{n,p}))$ represent large-dimensional counterparts of the finite-sample optimal eigenvalues $\boldsymbol{d}_n^* = (d_{n,1}^*, \dots, d_{n,p}^*)$ of equation (2.2). $\boldsymbol{d}_n^{\mathrm{o}}$ is an *oracle* estimator, meaning that it cannot be computed from observable data, since it depends on the limiting sample spectral density f, its Hilbert transform $\mathcal{H}_f$, and the limiting concentration ratio c. Nonetheless, it constitutes a useful stepping stone toward the ultimate objective, which is the construction of a *bona fide* estimator (i.e., one that is feasible in practice) with the same asymptotic properties.

REMARK 3.4. Ledoit and Wolf ((2018), Section 4.2) prove that the estimator S_n^{o} is also optimal within the class of rotation-equivariant estimators of Assumption 3.4 with respect to the Frobenius loss. Intuitively, this is because the two corresponding finite-sample optimal estimators are identical, as pointed out in Remark 2.2.

There is considerable interest in estimating the nonlinearly shrunk eigenvalues $\boldsymbol{d}_n^{\mathrm{o}}$ from $\boldsymbol{\lambda}_n$ only. For the limiting concentration ratio c, there is no problem: we can just plug its natural estimator $c_n := p/n$ into formula (3.3). Things are more complicated, however, for the limiting sample spectral density f and its Hilbert transform $\mathcal{H}_f$. Given that the sample spectral e.d.f. F_n converges to F almost surely, the obvious idea would have been to plug its derivative F_n' in place of f:

$$\frac{\lambda_{n,i}}{[\pi \frac{p}{n}\lambda_{n,i}F_n'(\lambda_{n,i})]^2 + [1 - \frac{p}{n} - \pi\frac{p}{n}\lambda_{n,i}\mathcal{H}_{F_n'}(\lambda_{n,i})]^2}. \tag{3.6}$$

Unfortunately, this approach cannot work because F_n is discontinuous at every $\lambda_{n,i}$, so its derivative does not exist at these points, and *a fortiori* the Hilbert transform of F_n' does not exist either. This problem has been a major stumbling block in the literature. The new solution that we propose is to use kernel estimators to estimate f and $\mathcal{H}_f$.

4. Asymptotic theory.

4.1. *Kernel requirements.*

ASSUMPTION 4.1 (Kernel). Let $k(x)$ denote a continuous, symmetric, nonnegative probability density function (p.d.f.) whose support is a compact interval $[-R, R]$, with mean zero and variance one. We assume throughout that this kernel satisfies the following conditions:

1. Its Hilbert transform $\mathcal{H}_k$ exists and is continuous on $\mathbb{R}$.
2. Both the kernel k and its Hilbert transform $\mathcal{H}_k$ are functions of bounded variation.

4.2. *Proportional bandwidth.* The approach that we propose uses a variable bandwidth proportional to the magnitude of a given sample eigenvalue. Thus, the bandwidth applied to the sample eigenvalue $\lambda_{n,i}$ is equal to $h_{n,i} := \lambda_{n,i} h_n$, for $i = 1, \dots, p$, where h_n is a vanishing sequence of positive numbers to be specified below. In terms of nomenclature, we can call h_n the "global bandwidth" and $h_{n,i}$ a "locally adaptive" bandwidth.

The advantages of the proportional bandwidth relative to the simpler and more common fixed one are threefold. First, if $h_n < 1/R$, which will be the case for large enough n, then the support of the kernel estimator will remain in the positive half of the real line. This is desirable because the covariance matrix is positive definite. Second, estimating a covariance matrix is a scale-equivariant problem: If we multiply all the variables by some $\alpha \neq 0$, then the estimator should remain exactly the same except for rescaling by the coefficient α^2. Using a global bandwidth that depends solely on n but not on the scale of the eigenvalues would violate this desirable feature. Third, the mathematical nature of the mapping $(c, H) \mapsto F_{c,H}$ is such that large eigenvalues get smudged more than small ones. Given the somewhat qualitative nature of this statement, a visual illustration shall suffice.

In Figure 3, the small eigenvalues (to the left) get spread out less than the large ones (to the right). Indeed, the width of the support interval associated with a given eigenvalue is almost exactly proportional to the magnitude of the eigenvalue itself. This is why a "one-size-fits-all" approach to bandwidth selection is ill-suited for the estimation of the spectral density.

Additional justification for proportional bandwidth is given by the "arrow model" of Ledoit and Wolf (2018). This model shows that, if the largest population eigenvalue $\tau_{n,p}$ becomes very large and detaches itself from the bulk of the other population eigenvalues, then the corresponding sample eigenvalue will also detach itself, and fall somewhere within an interval of width proportional to $\tau_{n,p}$.

A similar phenomenon occurs in the simple case where all but one of the population eigenvalues are equal to zero. Then all sample eigenvalues but one are equal to zero, and the nonzero eigenvalue behaves like a variance. It is well known that the standard deviation of the sample variance (based on i.i.d. data) is proportional to the population variance. Under traditional (finite-dimensional) asymptotics, it has been long known since Girshick ((1939), page 217) that the limiting standard deviation of a sample eigenvalue is directly proportional to the eigenvalue itself.

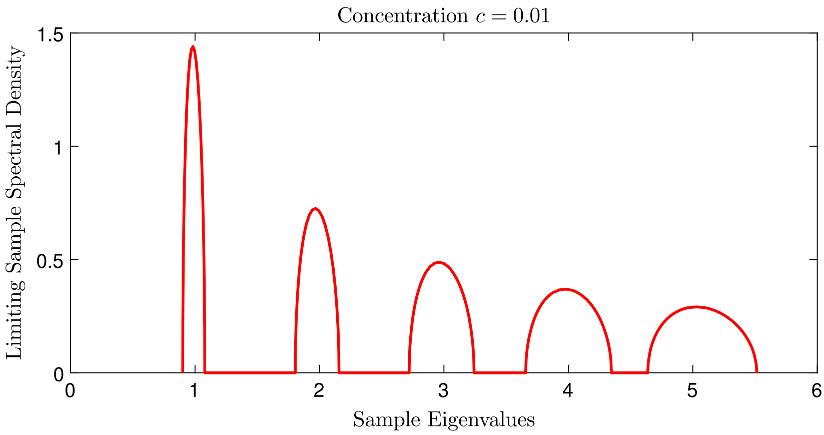

FIG. 3. *Limiting sample spectral density f when the population eigenvalues are $\{1, 2, \dots, 5\}$, each with weight* $1/5$.

4.3. *Kernel estimators.* The kernel estimator of the sample spectral p.d.f. f is

$$\forall x \in \mathbb{R} \quad \widetilde{f}_n(x) := \frac{1}{p}\sum_{i=1}^{p}\frac{1}{h_{n,i}}k\left(\frac{x-\lambda_{n,i}}{h_{n,i}}\right) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{\lambda_{n,i}h_n}k\left(\frac{x-\lambda_{n,i}}{\lambda_{n,i}h_n}\right).$$

The kernel estimator of its Hilbert transform $\mathcal{H}_f$ is

$$\mathcal{H}_{\widetilde{f}_n}(x) := \frac{1}{p}\sum_{i=1}^{p}\frac{1}{h_{n,i}}\mathcal{H}_k\left(\frac{x-\lambda_{n,i}}{h_{n,i}}\right) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{\lambda_{n,i}h_n}\mathcal{H}_k\left(\frac{x-\lambda_{n,i}}{\lambda_{n,i}h_n}\right) = \frac{1}{\pi}\,\mathrm{PV}\int\frac{\widetilde{f}_n(t)}{x-t}\,dt.$$

4.4. *Uniform consistency.* Our main results are as follows. All proofs are in Appendix A in the Supplementary Material (Ledoit and Wolf (2020)).

THEOREM 4.1. *Suppose that the kernel $k(x)$ satisfies the conditions of Section 4.1. Let h_n be a sequence of positive numbers satisfying*

$$\lim_{n\to\infty} nh_n^{5/2} = \infty \quad \textit{and} \quad \lim_{n\to\infty} h_n = 0. \tag{4.1}$$

Moreover, suppose that Assumptions 3.1–3.3 are satisfied. Then, both

$$\sup_{x\in\mathsf{Supp}(F)} \left|\widetilde{f}_n(x) - f(x)\right| \longrightarrow 0 \quad \textit{and} \quad \sup_{x\in\mathsf{Supp}(F)} \left|\mathcal{H}_{\widetilde{f}_n}(x) - \mathcal{H}_f(x)\right| \longrightarrow 0 \tag{4.2}$$

in probability.

The two kernel estimators enable us to shrink the sample eigenvalues nonlinearly as follows:

$$\forall i = 1, \dots, p \quad \widetilde{d}_{n,i} := \frac{\lambda_{n,i}}{[\pi\frac{p}{n}\lambda_{n,i}\widetilde{f}_n(\lambda_{n,i})]^2 + [1-\frac{p}{n}-\pi\frac{p}{n}\lambda_{n,i}\mathcal{H}_{\widetilde{f}_n}(\lambda_{n,i})]^2}. \tag{4.3}$$

The shrunk eigenvalues $\widetilde{\boldsymbol{d}}_n := (\widetilde{d}_{n,1}, \dots, \widetilde{d}_{n,p})'$ are then stacked into the diagonal of the diagonal matrix $\widetilde{D}_n$ to generate a covariance matrix estimator

$$\widetilde{S}_n := U_n\widetilde{D}_nU_n' = \sum_{i=1}^{p}\widetilde{d}_{n,i}\cdot u_{n,i}u_{n,i}'. \tag{4.4}$$

The covariance matrix estimator based on the kernel method performs as well in the large-dimensional asymptotic limit as the nonlinear shrinkage estimator of Ledoit and Wolf (2012, 2015) based on the indirect method, as the following corollary attests.

COROLLARY 4.1. *Under Assumptions 3.1–3.4, the covariance matrix estimator $\widetilde{S}_n$ minimizes in the class of nonlinear shrinkage estimators defined in Assumption 3.4 the limit in probability of the minimum variance loss function $\mathcal{L}_n^{\mathrm{MV}}$, as p and n go to infinity together.*

REMARK 4.1. The above statement remains true if the minimum variance loss $\mathcal{L}_n^{\mathrm{MV}}$ is replaced with the Frobenius loss $\mathcal{L}_n^{\mathrm{FR}}$ instead. Indeed, Section 4.2 of Ledoit and Wolf (2018) proves that the same estimator S_n^{o} is also optimal within the class of rotation-equivariant estimators of Assumption 3.4 with respect to the Frobenius loss.

4.5. *Epanechnikov kernel.* The two most popular kernels for density estimation are the Gaussian kernel and the Epanechnikov (1969) kernel. We choose the latter for four reasons:

Common sense. The support of the Gaussian kernel is the real line, yet covariance matrix eigenvalues cannot be negative. By contrast, the Epanechnikov kernel has bounded support.

Asymptotic theory. Assumption 4.1 requires a kernel with bounded support for uniform consistency to hold as per Theorem 4.1.

Optimality. The Epanechnikov kernel minimizes mean-squared error, at least for i.i.d. data.

Computation. The Hilbert transform of the Gaussian kernel is the Dawson (1898) integral, which is a higher-transcendental function extremely slow to compute.

The kernel originally introduced by Epanechnikov ((1969), equation (13)) has unit variance, support $[-\sqrt{5}, \sqrt{5}]$, and density

$$\forall x \in \mathbb{R} \quad \kappa^E(x) := \frac{3}{4\sqrt{5}}\left[1 - \frac{x^2}{5}\right]^+, \tag{4.5}$$

where $[\,\cdot\,]^+$ denotes the positive part of a real number. This is the default kernel used for univariate density estimation by the popular software STATA, among others. Its Hilbert transform does not appear to have been computed in the literature. We derive it in the following proposition.

PROPOSITION 4.1.

$$\forall x \in \mathbb{R} \quad \mathcal{H}_{\kappa^E}(x) = \begin{cases} -\dfrac{3x}{10\pi} + \dfrac{3}{4\sqrt{5}\pi}\left(1 - \dfrac{x^2}{5}\right)\log\left|\dfrac{\sqrt{5}-x}{\sqrt{5}+x}\right| & \textit{if } |x| \neq \sqrt{5}, \\ -\dfrac{3x}{10\pi} & \textit{if } |x| = \sqrt{5}. \end{cases} \tag{4.6}$$

PROPOSITION 4.2. *The Epanechnikov kernel satisfies Assumption 4.1.*

From this, we deduce for all $i = 1, \dots, p$

$$\widetilde{f}_n(\lambda_{n,i}) = \frac{1}{p}\sum_{j=1}^{p} \frac{3}{4\sqrt{5}\lambda_{n,j}h_n}\left[1 - \frac{1}{5}\left(\frac{\lambda_{n,i}-\lambda_{n,j}}{\lambda_{n,j}h_n}\right)^2\right]^+, \tag{4.7}$$

$$\begin{aligned} \mathcal{H}_{\widetilde{f}_n}(\lambda_{n,i}) = \frac{1}{p}\sum_{j=1}^{p}\Bigg\{ & -\frac{3(\lambda_{n,i}-\lambda_{n,j})}{10\pi\lambda_{n,j}^2 h_n^2} + \frac{3}{4\sqrt{5}\pi\lambda_{n,j}h_n}\left[1 - \frac{1}{5}\left(\frac{\lambda_{n,i}-\lambda_{n,j}}{\lambda_{n,j}h_n}\right)^2\right] \\ & \times \log\left|\frac{\sqrt{5}\lambda_{n,j}h_n - \lambda_{n,i} + \lambda_{n,j}}{\sqrt{5}\lambda_{n,j}h_n + \lambda_{n,i} - \lambda_{n,j}}\right|\Bigg\}. \end{aligned} \tag{4.8}$$

Throughout, the last term in (4.8) is understood to be a zero in the unlikely event that $(\lambda_{n,i} - \lambda_{n,j})^2$ happens to be exactly equal to $5\lambda_{n,j}^2 h_n^2$. Alternative kernels are explored as robustness checks through Monte Carlo simulations in Appendix D in the Supplementary Material (Ledoit and Wolf (2020)).

4.6. *Choice of bandwidth.* The most consequential choice relating to the bandwidth has already been justified in Section 4.2: We introduced a locally adaptive bandwidth proportional to the magnitude of the sample eigenvalues: $h_{n,i} = \lambda_{n,i}h_n$.

As for the global bandwidth h_n, Theorem 4.1 requires it to be a negative exponent of n strictly less than $2/5$. Jing et al. (2010) is the only previous article we are aware of that uses a kernel to estimate the limiting sample spectral density. They select the exponent $1/3$, which is the first "simple" fraction below the authorized boundary of $2/5$. There is always the potential for disagreement about what the "right" exponent should be, so to anchor on solid ground we just follow in their footsteps:

$$h_n := n^{-1/3} \quad \Longrightarrow \quad \forall i = 1, \dots, p \quad h_{n,i} := \lambda_{n,i} h_n = \lambda_{n,i} n^{-1/3}. \tag{4.9}$$

The kernel, location- and bandwidth-adjusted for the ith sample eigenvalue,

$$\frac{1}{h_{n,i}} \kappa^E\left(\frac{x - \lambda_{n,i}}{h_{n,i}}\right),$$

has support is $[\lambda_{n,i}(1 - \sqrt{5}n^{-1/3}), \lambda_{n,i}(1 + \sqrt{5}n^{-1/3})]$. The lower boundary is in the positive half-line if and only if $n > 5\sqrt{5} \approx 11.2$, therefore it is unadvisable to use this large-dimensional asymptotic procedure when $p < 12$. Alternative choices of the bandwidth are explored as robustness checks through Monte Carlo simulations in Appendix D in the Supplementary Material (Ledoit and Wolf (2020)).

4.7. *Summary*: *The analytical nonlinear shrinkage estimator.* Compute the spectral decomposition of the sample covariance matrix as per Section 2.1:

$$S_n =: \sum_{i=1}^{p} \lambda_{n,i} \cdot u_{n,i} u'_{n.i}.$$

Choose the global bandwidth as per Section 4.6:

$$h_n := n^{-1/3}.$$

Specify the locally adaptive bandwidth as per Section 4.2:

$$\forall j = 1, \dots, n \quad h_{n,j} := \lambda_{n,j} h_n.$$

Estimate the spectral density with the Epanechnikov kernel from Section 4.5:

$$\widetilde{f}_n(\lambda_{n,i}) := \frac{1}{p} \sum_{j=1}^{p} \frac{3}{4\sqrt{5} h_{n,j}} \left[1 - \frac{1}{5}\left(\frac{\lambda_{n,i} - \lambda_{n,j}}{h_{n,j}}\right)^2\right]^+,$$

and its Hilbert transform as per Section 3.3 and Proposition 4.1:

$$\mathcal{H}_{\widetilde{f}_n}(\lambda_{n,i}) := \frac{1}{p} \sum_{j=1}^{p} \left\{ -\frac{3(\lambda_{n,i} - \lambda_{n,j})}{10\pi h_{n,j}^2} + \frac{3}{4\sqrt{5}\pi h_{n,j}} \left[1 - \frac{1}{5}\left(\frac{\lambda_{n,i} - \lambda_{n,j}}{h_{n,j}}\right)^2\right] \times \log\left|\frac{\sqrt{5}h_{n,j} - \lambda_{n,i} + \lambda_{n,j}}{\sqrt{5}h_{n,j} + \lambda_{n,i} - \lambda_{n,j}}\right| \right\}.$$

Compute the asymptotically optimal nonlinear shrinkage formula as per Section 3.4:

$$\forall i = 1, \dots, p \quad \widetilde{d}_{n,i} := \frac{\lambda_{n,i}}{[\pi \frac{p}{n} \lambda_{n,i} \widetilde{f}_n(\lambda_{n,i})]^2 + [1 - \frac{p}{n} - \pi \frac{p}{n} \lambda_{n,i} \mathcal{H}_{\widetilde{f}_n}(\lambda_{n,i})]^2}.$$

Recompose the covariance matrix estimator as per Section 4.4:

$$\widetilde{S}_n := \sum_{i=1}^{p} \widetilde{d}_{n,i} \cdot u_{n,i} u'_{n,i}.$$

Computer code for this analytical estimator is in Appendix E in the Supplementary Material (Ledoit and Wolf (2020)).

5. Monte Carlo simulations.

5.1. *Competitors*. We compare the performance of the following six covariance matrix estimators:

Sample. The sample covariance matrix S_n.

Linear. The linear shrinkage estimator of Ledoit and Wolf (2004).

Analytical. The analytical nonlinear shrinkage estimator $\widetilde{S}_n$ of Section 4.7.

QuEST. The nonlinear shrinkage estimator of Ledoit and Wolf (2015), which is based on numerical inversion of the QuEST function.

NERCOME. The Nonparametric Eigenvalue-Regularized COvariance Matrix Estimator of Lam (2016), which is based on sample splitting.

FSOPT. The finite-sample optimal estimator S_n^* defined in equation (2.3) which requires knowledge of the population covariance matrix Σ_n; therefore, this estimator is not feasible in the real world but it serves as a useful benchmark in Monte Carlo simulations.

The Matlab code for NERCOME was generously provided by Professor Clifford Lam from the Department of Statistics at the London School of Economics. The code for the QuEST package comes from the numerical implementation detailed in Ledoit and Wolf (2017b) and is freely downloadable from the academic website of the second author.

REMARK 5.1. Corollary 4.1 implies that, under large-dimensional asymptotics, Analytical and QuEST have the same limiting loss. Therefore, we should expect their performances to be nearly identical for large (p, n). For small and moderate (p, n), we would expect the performance of QuEST to be somewhat better because it exploits the feature that f is the density of a limiting sample spectral distribution F, which is an output of the fundamental equation of random matrix theory; hence, QuEST can be considered a model-based estimator. By contrast, Analytical does not exploit this feature of f, and thus can be considered model-free.

5.2. *Percentage relative improvement in average loss*. The main quantity of interest is the Percentage Relative Improvement in Average Loss (PRIAL). It is defined for a generic estimator $\widehat{\Sigma}_n$ as

$$\text{PRIAL}_n^{\text{MV}}(\widehat{\Sigma}_n) := \frac{\mathbb{E}[\mathcal{L}_n^{\text{MV}}(S_n, \Sigma_n)] - \mathbb{E}[\mathcal{L}_n^{\text{MV}}(\widehat{\Sigma}_n, \Sigma_n)]}{\mathbb{E}[\mathcal{L}_n^{\text{MV}}(S_n, \Sigma_n)] - \mathbb{E}[\mathcal{L}_n^{\text{MV}}(S_n^*, \Sigma_n)]} \times 100\%, \tag{5.1}$$

where $\mathcal{L}_n^{\text{MV}}$ denotes the minimum-variance loss function of Section 2.2, Σ_n denotes the population covariance matrix, and S_n^* denotes the finite-sample-optimal rotation-equivariant estimator of equation 2.3, which is only observable in Monte Carlo simulations but not in reality. The expectation $\mathbb{E}[\cdot]$ is in practice taken as the average across $\max\{100, \min\{1000, 10^5/p\}\}$ Monte Carlo simulations; for example, in dimension $p = 500$, we run 200 simulations instead of 1000. We do so because in higher dimensions the results are more stable across random simulations, so it is not necessary to run so many.

By construction, $\text{PRIAL}_n^{\text{MV}}(S_n) = 0\%$. It means that the sample covariance matrix represents the baseline reference against which any loss reduction is measured. An estimator that has lower (higher) expected loss than the sample covariance matrix will score a positive (negative) PRIAL.

Also by construction, $\text{PRIAL}_n^{\text{MV}}(S_n^*) = 100\%$ because this is the maximum amount of loss reduction that can be attained by nonlinear shrinkage within the rotation-equivariant framework of Section 2.1, as shown in Proposition 2.1. Given that the construction of S_n^* requires knowledge of the population covariance matrix, 100% improvement represents an upper limit that is unattainable in reality. The question is how close to the gold standard of 100% a *bona fide* estimator can get.

Recall that the loss function $\mathcal{L}_n^{\mathrm{MV}}$ represents the true variance of the linear combination of the original variables that has minimum estimated variance under generic linear constraint, suitably normalized. Therefore, the PRIAL measures how much of the potential for variance reduction is captured by any given shrinkage technique.

5.3. *Baseline scenario.* The simulations are organized around a baseline scenario, where each parameter will be subsequently varied in order to assess the robustness of the conclusions. The baseline scenario has the following characteristics:

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

The distribution of the population eigenvalues is a particularly interesting and difficult case introduced and analyzed in detail by Bai and Silverstein (1998).

Table 2 reports estimator performances under the baseline scenario. Computational times in milliseconds come from a 64-bit, quad-core 4.00 GHz Windows desktop PC running Matlab R2016a.

The 0% PRIAL for the sample covariance matrix and the 100% PRIAL for the finite-sample optimal estimator are by construction. Linear shrinkage captures half of the potential for variance reduction. Nonlinear shrinkage captures 92%–98% of the potential, depending on the method used (NERCOME/Analytical/QuEST), which is a very satisfactory number.

One key lesson is that the analytical estimator developed in the present paper is faster than the other two nonlinear shrinkage estimators by two orders of magnitude. Thus, it delivers the best of both worlds: QuEST-tier variance reduction at Linear-tier speed. Note also that 2 of the 3 milliseconds spent on computing the analytical formula are spent on extracting the eigenvalues and eigenvectors of the sample covariance matrix, an operation that all nonlinear shrinkage estimator must perform, even if they use knowledge of the population covariance matrix (cf. FSOPT).

The only estimator that is in the same ballpark as analytical nonlinear shrinkage in terms of both speed and accuracy is the finite-sample optimal estimator, which is not feasible in practice. Among *bona fide* estimators, the analytical nonlinear estimator is the only one that comes even close to matching both the speed and accuracy of the finite-sample optimal estimator.

Table 2 demonstrates that applied researchers who are already comfortable with linear shrinkage and would like to upgrade to nonlinear shrinkage for performance enhancement,

TABLE 2
Simulation results for the baseline scenario

Estimator	Sample	Linear	Analytical	QuEST	NERCOME	FSOPT
Average Loss	2.71	2.10	1.52	1.50	1.58	1.48
PRIAL	0%	50%	97%	98%	92%	100%
Time (ms)	1	2	3	2346	3071	3

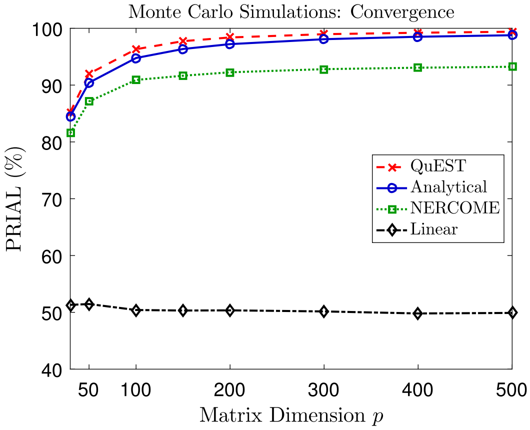

FIG. 4. *Evolution of the PRIAL of various estimators as the matrix dimension and the sample size go to infinity together.*

but have been concerned by the numerical complexity of the earlier techniques, can now safely adopt the analytical estimator.

5.4. *Convergence.*

5.4.1. *Large-dimensional asymptotic performance.* Under large-dimensional asymptotics, the matrix dimension p and the sample size n go to infinity together, while their ratio p/n converges to some limit c. In the first experiment, we let p and n vary together, with their ratio fixed at the baseline value of $1/3$. The results are displayed in Figure 4.

The three nonlinear shrinkage estimators perform approximately the same as one another. They do well even in small dimensions, but do better as the dimension grows large. The difference between the PRIALs of QuEST and Analytical is never more than 2%, which is very small.

5.4.2. *Speed.* Apart from minimizing the asymptotic loss, a key advantage of the analytical estimator proposed in the present paper is that it is fast regardless of the matrix dimension. The computation times needed to produce Figure 4 are displayed in Figure 5.

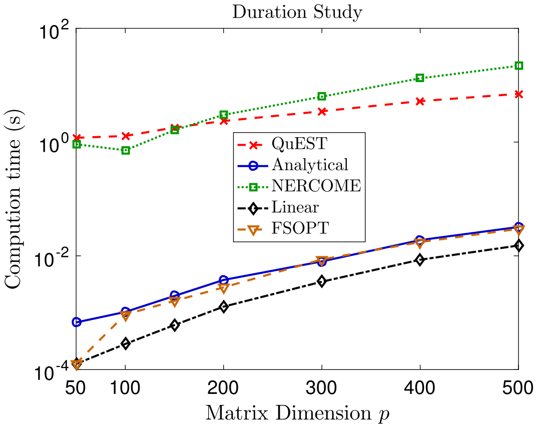

FIG. 5. *Computational speed of various shrinkage estimators as the matrix dimension and the sample size go to infinity together, measured in seconds, with log-scale on the vertical axis.*

TABLE 3

Result of 100 *Monte Carlo simulations for dimension* $p = 10{,}000$ *and sample size* $n = 30{,}000$

Estimator	Sample	Linear	Analytical	FSOPT
Average Loss	2.679	2.086	1.488	1.487
PRIAL	0%	49.74%	99.90%	100%
Time (s)	21	43	113	108

There is a clear gap between, on the one hand, QuEST and NERCOME and, on the other hand, Analytical, Linear and FSOPT. Analytical nonlinear shrinkage is typically 1000 times faster than its numerical counterparts.

5.4.3. *Ultra-large dimension.* The analytical formula enables us to apply nonlinear shrinkage in much larger dimensions than was previously imaginable within reasonable time. To prove the point, we reproduce Table 2 for 50-times larger dimension and sample size, with the fast estimators only. The results are presented in Table 3.

The first item of note is that the PRIAL of the analytical nonlinear shrinkage estimator gets ever closer to 100%, as expected from theory.

Speed-wise, it takes less than two minutes to compute the analytical nonlinear shrinkage formula in dimension 10,000. Most of the time is spent computing the sample covariance matrix ($O(p^2n)$ computational cost), extracting its eigenvalues and eigenvectors ($O(p^3)$ cost), and recombining the sample eigenvectors with the shrunk eigenvalues as per (4.4) (also $O(p^3)$ cost). These operations would be necessary for any nonlinear shrinkage estimator—even if one used knowledge of the population covariance matrix, as evidenced by the FSOPT speed in the right-most column. The actual computation of the kernel estimator of the Hilbert transform $\mathcal{H}_{\tilde{f}_n}$ as defined in Section 4.3 and of the shrunk eigenvalues themselves (4.3), which are the only steps specific to this method as opposed to any other nonlinear shrinkage, just take 4 seconds in total because they require one order of magnitude fewer floating point operations: only $O(p^2)$.

Further (unreported) simulations in dimension $p = 20{,}000$ with sample size $n = 60{,}000$ show computation times 7.6 to 8.9 times longer for the four estimators of Table 3, which tightly brackets the theoretical prediction of $2^3 = 8$ based on the reasoning of the previous paragraph.

5.5. *Concentration ratio.* We vary the concentration ratio p/n from 0.1 to 0.9 while holding the product $p \times n$ constant at the level it had under the baseline scenario, namely, $p \times n = 120{,}000$. The PRIALs are displayed in Figure 6.

Linear shrinkage performs very well in high concentrations but does not beat the sample covariance matrix for low concentrations. Appendix B.1 in the Supplementary Material (Ledoit and Wolf (2020)) shows that this finding is due to the fact that linear shrinkage is optimized for a different loss function than the minimum variance loss, namely, the Frobenius loss. Under Frobenius loss, linear shrinkage always beats the sample covariance matrix in the same simulation experiment.

The three nonlinear shrinkage estimators perform approximately the same as one another, with Analytical in particular being very close to QuEST and above the 96% mark across the board.

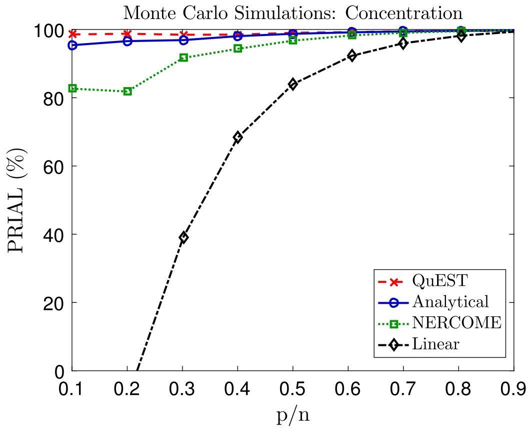

FIG. 6. *Evolution of the PRIAL of various estimators as a function of the ratio of the matrix dimension to the sample size.*

5.6. *Condition number.* We start again from the baseline scenario and, this time, vary the condition number θ of the population covariance matrix. We set 20% of the population eigenvalues equal to 1, 40% equal to $(2\theta + 7)/9$ and 40% equal to θ. Thus, the baseline scenario corresponds to $\theta = 10$. In this experiment, we let θ vary from $\theta = 3$ to $\theta = 30$. This corresponds to linearly squeezing or stretching the distribution of population eigenvalues. The resulting PRIALs are displayed in Figure 7.

Linear shrinkage performs very well for low condition numbers, but not so well for high condition numbers; once again, one must bear in mind that this is due to the fact that it is optimized for a different loss function that the one we use here. Appendix B.2 in the Supplementary Material (Ledoit and Wolf (2020)) verifies this by running the same simulations again under Frobenius loss and showing that linear shrinkage dominates the sample covariance matrix across the board in this metric.

The three nonlinear shrinkage estimators all capture a very high percentage of the potential for variance reduction, with Analytical in particular being very close to QuEST and above the 97% mark across the board.

5.7. *Nonnormality.* In this experiment, we start from the baseline scenario and change the distribution of the variates. We study the Bernoulli coin-toss distribution, which is the

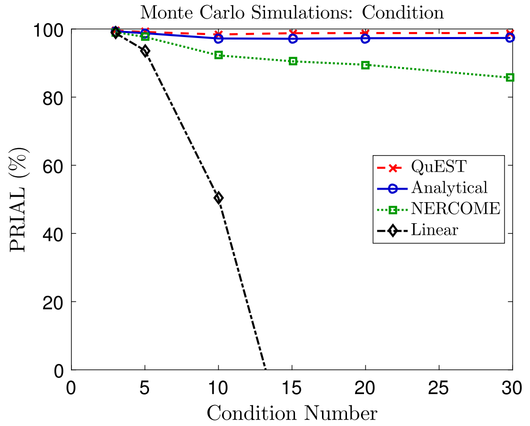

FIG. 7. *Evolution of the PRIAL of various estimators as a function of the condition number of the population covariance matrix.*

TABLE 4
Simulation results for various variate distributions (PRIAL)

Distribution	Linear	Analytical	QuEST	NERCOME
Bernoulli	51%	97%	98%	92%
Laplace	50%	97%	98%	92%
‘Student’ t_5	49%	97%	98%	92%

most platykurtic of all distributions, the Laplace distribution, which is leptokurtotic, and the “Student” t-distribution with 5 degrees of freedom, also leptokurtotic. All of these are suitably normalized to have mean zero and variance one, if necessary. The results are presented in Table 4.

This experiment confirms that the results of the baseline scenario are not sensitive to the distribution of the variates.

5.8. *Shape of the distribution of population eigenvalues.* Relative to the baseline scenario, we now move away from the clustered distribution for the population eigenvalues and try a variety of continuous distributions drawn from the Beta family. They are linearly shifted and stretched so that the support is $[1, 10]$. A graphical illustration of the densities of the various Beta shapes studied below can be found in Ledoit and Wolf ((2012), Figure 7). The results are presented in Table 5.

Note that the 100% PRIALs are due to rounding effect: no PRIAL ever exceeds 99.8%. This time, linear shrinkage does much better overall, except perhaps for the bimodal shape $(0.5, 0.5)$. This is due to the fact that, in the seven other cases, the optimal nonlinear shrinkage formula happens to be almost linear. The three nonlinear shrinkage estimators capture a very high percentage of the potential for variance reduction in all cases, with Analytical being virtually indistinguishable from QuEST and above the 97% mark across the board.

5.9. *Fixed-dimensional asymptotics.* An instructive experiment that falls outside the purview of large-dimensional asymptotics is to keep the dimension p fixed at the level specified by the baseline scenario, while letting the sample size n go to infinity. This is standard, or fixed-dimensional, asymptotics. We let the sample size grow from $n = 250$ to $n = 20{,}000$. The results are displayed in Figure 8.

Linear shrinkage performs well for small sample sizes but not for large ones. This is to be expected given Figure 6 because small (large) sample sizes correspond to large (small)

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

Beta parameters	Linear	Analytical	QuEST	NERCOME
(1, 1)	83%	98%	99%	96%
(1, 2)	95%	99%	99%	98%
(2, 1)	94%	99%	99%	99%
(1.5, 1.5)	92%	99%	99%	98%
(0.5, 0.5)	50%	98%	98%	94%
(5, 5)	98%	100%	100%	99%
(5, 2)	97%	100%	100%	98%
(2, 5)	99%	99%	99%	99%

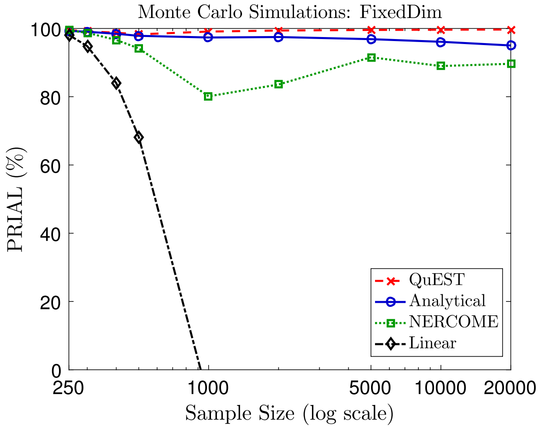

FIG. 8. *Evolution of the PRIAL as the sample size grows toward infinity, while the matrix dimension remains fixed.*

concentration ratios. Appendix B.3 in the Supplementary Material (Ledoit and Wolf (2020)) shows that linear shrinkage does not suffer from any such weakness under the Frobenius loss.

The three nonlinear shrinkage estimators all capture a very high percentage of the potential for variance reduction, with Analytical in particular being very close to QuEST and above the 96% mark across the board.

5.10. *Arrow model.* A standard assumption under large-dimensional asymptotics is that the largest population eigenvalue remains uniformly bounded even as the dimension goes to infinity. However, in the real world, it is possible to encounter a pervasive factor that generates an eigenvalue of the same order of magnitude as p. Therefore, it is useful to see how shrinkage would perform under such a violation of the original assumptions.

Inspired by a factor model where all pairs of variables have 50% correlation and all variables have unit standard deviation, and by the "arrow model" introduced by Ledoit and Wolf ((2018), Section 7), we set the largest eigenvalue (the "arrow") equal to $1 + 0.5(p - 1)$. The other eigenvalues (the "bulk") are drawn from the left-skewed Beta(5, 2) distribution, shifted and stretched linearly so that it has support [1, 10] (cf. row 7 of Table 5). The results are displayed in Figure 9, where the matrix dimension varies from $p = 50$ to $p = 500$.

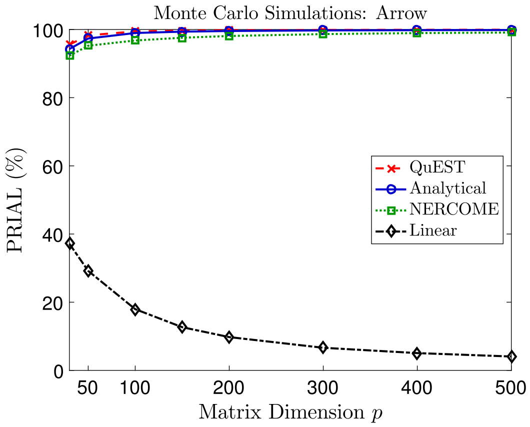

FIG. 9. *Evolution of the PRIAL as the matrix dimension, the top eigenvalue and the sample size all go to infinity together.*

Linear shrinkage improves upon the sample covariance matrix, but it overshrinks the arrow eigenvalue and undershrinks the bulk. The three nonlinear shrinkage estimators do not have this problem; in particular, Analytical is always above the 94% mark.

5.11. *Summary.* The results of this extensive set of Monte Carlo simulations are very consistent. Linear shrinkage does a good job in most cases, and in some cases an excellent one. Appendix B in the Supplementary Material (Ledoit and Wolf (2020)) shows that any instance of below-par performance is solely due to the "unfair" choice of a loss criterion with respect to which it was not optimized.

The three nonlinear shrinkage estimators perform very well across the board. Their performance levels are roughly similar to one another and of high standard. If anything, QuEST tends to be better than Analytical, which tends to be better than NERCOME, but the differences are relatively small, and there are exceptions. Between QuEST and Analytical there is hardly any difference at all. These findings are consistent with our previous expectations voiced in Remark 5.1; it is reassuring to see that even for small and moderate (p, n), Analytical performs almost as well as QuEST.

The analytical nonlinear shrinkage estimator is very simple to implement, as proven by the 20-line Matlab code in Appendix E in the Supplementary Material (Ledoit and Wolf (2020)). It captures 90% or more of the potential for variance reduction that comes from shrinking the sample eigenvalues. It is typically 1000 times faster than the other nonlinear shrinkage estimators, and is the only one that can handle ultralarge dimensions up to 10,000 and more in reasonable time.

5.12. *Robustness checks.* Appendix D in the Supplementary Material (Ledoit and Wolf (2020)) presents extensive robustness checks that examine the extent to which the performance of the analytical nonlinear shrinkage estimator is sensitive to the choices of kernel and bandwidth.

This thorough investigation of potential variants to the analytical formula of Section 4.7 reveals that nothing of value can be gained from changing any of the specification choices in the kernel estimation part. Nothing is lost either by using the triangular or the semicircular kernel, or by varying the global bandwidth exponent α in the reasonable range of $[0.2, 0.35]$. These findings show that our approach is robust and that its value lies not in some happenstance specification of kernel and bandwidth, but in correctly identifying and mathematically exploiting the deep connection between kernel estimation of the sample spectral density and optimal nonlinear shrinkage of large-dimensional covariance matrices through the Hilbert transform.

6. Conclusion.

This paper develops the first analytical formula for asymptotically optimal nonlinear shrinkage of large-dimensional covariance matrices. The formula was derived by introducing kernel estimation, not only of the density itself (which has been done before), but also of its Hilbert transform as a worthy mathematical procedure. A density and its Hilbert transform are 'joined at the hip' in the sense that they are, respectively, the imaginary and real part of the unique analytic extension of a real function into the complex plane. Venturing into the complex plane is a technical necessity not only for large-dimensional random matrix theory but also for signal processing, among other fields.

Another innovation is to estimate the two ingredients in the optimal nonlinear shrinkage formula, namely, the limiting sample spectral density and its Hilbert transform, with a proportional-bandwidth kernel estimator reflective of the scale-equivariance of the problem. The resulting computations are analytical in nature, easy to understand, straightforward to implement, fast and scalable.

Extensive Monte Carlo simulations show that the analytical nonlinear shrinkage estimator captures a very high percentage (typically 96%+) of the potential for variance reduction that opens up when we shrink the eigenvalues of the sample covariance matrix. This means, in the context of finance, that one can design investment strategies that are as safe as they could possibly be, thus overcoming the "curse of dimensionality" which is often associated with portfolio selection involving large covariance matrices of stock returns.

The dimension of covariance matrices that can be handled successfully now is at least 10,000, one order of magnitude larger compared to the numerical nonlinear shrinkage estimators of Ledoit and Wolf (2015) and Lam (2016), which is an important bonus in the age of Big Data.

Acknowledgments. We would like to thank an Associate Editor and three anonymous referees for valuable comments which have resulted in an improved exposition of this paper.

SUPPLEMENTARY MATERIAL

Mathematical proofs and additional material (DOI: 10.1214/19-AOS1921SUPP; .pdf). This supplement contains detailed proofs of all mathematical results as well as additional material.

REFERENCES

ABADIR, K. M., DISTASO, W. and ŽIKEŠ, F. (2014). Design-free estimation of variance matrices. *J. Econometrics* **181** 165–180. MR3209862 https://doi.org/10.1016/j.jeconom.2014.03.010

BAI, Z. D. and SILVERSTEIN, J. W. (1998). No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *Ann. Probab.* **26** 316–345. MR1617051 https://doi.org/10.1214/aop/1022855421

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

DAWSON, H. G. (1898). On the Numerical Value of $\int_0^h e^{x^2}\,dx$. *Proc. Lond. Math. Soc.* **29** 519–522. MR1576451 https://doi.org/10.1112/plms/s1-29.1.519

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

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

EPANECHNIKOV, V. A. (1969). Nonparametric estimation of a multidimensional probability density. *Theory Probab. Appl.* **14** 153–158. MR0250422

ERDÉLYI, A., MAGNUS, W., OBERHETTINGER, F. and TRICOMI, F. G. (1954). *Tables of Integral Transforms. Vol. II.* McGraw-Hill Book Company, Inc., New York–Toronto–London. MR0065685

GIRSHICK, M. A. (1939). On the sampling theory of roots of determinantal equations. *Ann. Math. Stat.* **10** 203–224. MR0000127 https://doi.org/10.1214/aoms/1177732180

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

KRANTZ, S. G. (2009). *Explorations in Harmonic Analysis. Applied and Numerical Harmonic Analysis.* Birkhäuser, Inc., Boston, MA. MR2508404 https://doi.org/10.1007/978-0-8176-4669-1

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

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

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

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

LEDOIT, O. and WOLF, M. (2015). Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions. *J. Multivariate Anal.* **139** 360–384. MR3349498 https://doi.org/10.1016/j.jmva.2015.04.006

LEDOIT, O. and WOLF, M. (2017a). Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets Goldilocks. *Rev. Financ. Stud.* **30** 4349–4388.

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

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

LEDOIT, O. and WOLF, M. (2020). Supplement to "Analytical nonlinear shrinkage of large-dimensional covariance matrices." https://doi.org/10.1214/19-AOS1921SUPP.

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

MARKOWITZ, H. M. (1952). Portfolio selection. *J. Finance* **7** 77–91.

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

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

SILVERSTEIN, J. W. and BAI, Z. D. (1995). On the empirical distribution of eigenvalues of a class of large-dimensional random matrices. *J. Multivariate Anal.* **54** 175–192. MR1345534 https://doi.org/10.1006/jmva.1995.1051

SILVERSTEIN, J. W. and CHOI, S.-I. (1995). Analysis of the limiting spectral distribution of large-dimensional random matrices. *J. Multivariate Anal.* **54** 295–309. MR1345541 https://doi.org/10.1006/jmva.1995.1058

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

STIELTJES, T.-J. (1894). Recherches sur les fractions continues. *Ann. Fac. Sci. Univ. Toulouse Sci. Math. Sci. Phys.* **8** J1–J122. MR1508159