Ledoit, Olivier; Wolf, Michael

Working Paper

Direct nonlinear shrinkage estimation of large-dimensional covariance matrices

Working Paper, No. 264

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

Suggested Citation: Ledoit, Olivier; Wolf, Michael (2017) : Direct nonlinear shrinkage estimation of large-dimensional covariance matrices, Working Paper, No. 264, University of Zurich, Department of Economics, Zurich,
https://doi.org/10.5167/uzh-139880

This Version is available at:
https://hdl.handle.net/10419/173422

University of Zurich

Department of Economics

Working Paper Series

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

Working Paper No. 264

Direct Nonlinear Shrinkage Estimation of Large-Dimensional Covariance Matrices

Olivier Ledoit and Michael Wolf

September 2017

Direct Nonlinear Shrinkage Estimation of Large-Dimensional Covariance Matrices

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

September 2017

Abstract

This paper introduces a nonlinear shrinkage estimator of the covariance matrix that does not require recovering the population eigenvalues first. We estimate the sample spectral density and its Hilbert transform directly by smoothing the sample eigenvalues with a variable-bandwidth kernel. Relative to numerically inverting the so-called QuEST function, the main advantages of direct kernel estimation are: (1) it is much easier to comprehend because it is analogous to kernel density estimation; (2) it is only twenty lines of code in Matlab — as opposed to thousands — which makes it more verifiable and customizable; (3) it is 200 times faster without significant loss of accuracy; and (4) it can handle matrices of a dimension larger by a factor of ten. Even for dimension $10,000$, the code runs in less than two minutes on a desktop computer; this makes the power of nonlinear shrinkage as accessible to applied statisticians as the one of linear shrinkage.

KEY WORDS: Kernel estimation, Hilbert transform, large-dimensional asymptotics, nonlinear shrinkage, rotation equivariance.

JEL CLASSIFICATION NOS: C13.

1 Introduction

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

There exist two completely different nonlinear shrinkage methods that give satisfactory and largely compatible results. The first 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. Consistent results are achieved by numerical inversion of a deterministic multivariate function called the QuEST (acronym for Quantized Eigenvalues Sampling Transform) function, which essentially maps population eigenvalues into sample eigenvalues. The underlying framework is non-standard and based on *large-dimensional asymptotics* where the matrix dimension goes to infinity at the same rate as the sample size, with their ratio converging to some finite, nonzero limit called the limiting concentration ratio. 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 et al. (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 strictly be more accurate or faster, which it may not necessarily be (according to Monte Carlo simulations), but in the fact that it is decisively simpler and more transparent, thus providing an independent and easily verifiable sanity check for the mathematically delicate *indirect* method of the QuEST function.

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. This is achieved through nonparametric kernel estimation of the limiting spectral density of 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 only need twenty lines in Matlab); and from linear shrinkage we borrow the speed

and scalability.

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 towards 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. It is getting clear that, from the point of view of optimal covariance matrix estimation, the Hilbert transform is equally as important as the density itself. Krantz (2009) confirms that this is 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), we let it vary in proportion to the location of each 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 dimension attempted by nonlinear shrinkage before was $1,000$.

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 makes specific recommendations for the kernel function and the bandwidth parameter. Section 6 runs an extensive set of Monte Carlo simulations. Section 7 concludes.

2 Finite Samples

In this section, and this section only, the sample size n and covariance matrix dimension p are fixed and finite. This is 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 go to infinity under large-dimensional asymptotics.

2.1 Rotation Equivariance

Let Σ_n denote a p-dimensional population covariance matrix. A mean-zero 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 $\boldsymbol{\lambda}_n = (\lambda_{n,1}, \ldots, \lambda_{n,p})$ sorted in nondecreasing order without loss of generality, and an orthogonal matrix U_n whose columns $[u_{n,1} \ldots u_{n,p}]$ are 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 $\widehat{\boldsymbol{\delta}}_n = \left(\widehat{\delta}_{n,1}, \ldots, \widehat{\delta}_{n,p}\right) \in (0, +\infty)^p$ 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 dense in any other basis. Rotation equivariance does not take a stance on the orientation of the eigenvectors of the population covariance matrix.

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 et al., 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 et al. (2017, Definition 4.1). It can be expressed in our notation as

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

$\mathcal{L}_n^{\text{MV}}$ 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* $\mathcal{L}_n^{MV}$ *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* $\widehat{\delta}_{n,i} = \beta_n \cdot u_{n,i}' \Sigma_n u_{ni}$ *for* $i = 1, \ldots, p$.

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

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

prior to recombining it with the sample eigenvectors to form a covariance matrix estimator:

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

Remark 1. Section 3.1 of Ledoit and Wolf (2012) shows that the same estimator S_n^* is also optimal with respect to the (squared) Frobenius loss function, which is defined for generic estimator $\widehat{\Sigma}_n$ as

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

This is the loss function with respect to which Ledoit and Wolf's (2004) linear shrinkage estimator is optimized. Although $\mathcal{L}_n^{\text{FR}}$ does not constitute the main focus of the present paper, we take a look at it in Appendix B. ■

Whereas further investigations of the nonlinear shrinkage formula that maps $\boldsymbol{\lambda}_n$ into $\boldsymbol{d}_n^*$ are mathematically arduous or perhaps even unattainable in finite samples, decisive progress can be made by letting the dimension go to infinity.

3 Large-Dimensional Asymptotics

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 (2017c).

Assumption 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.

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

The e.d.f. returns the proportion of members of the collection that lie below its argument.

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

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

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

Assumption 3 (Data Generating Process)**.** X_n *is an* $n \times p$ *matrix of i.i.d. random variables with mean zero, variance one, and finite* 16*th moment. The matrix of observations is* $Y_n \coloneqq X_n \times \sqrt{\Sigma_n}$*. Neither* $\sqrt{\Sigma_n}$ *nor* X_n *are observed on their own: only* Y_n *is observed.*

Remark 2. 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 finite 4th moment is enough, and the Monte Carlo simulations we report in Table 4 appear to support this.

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

3.2 Random Matrix Theory

The literature on the eigenvalues of the sample covariance matrix under large-dimensional asymptotics is based on a foundational result by Marčenko and Pastur (1967). 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 1–3, there exists a limiting sample spectral distribution F such that $\forall x \in \mathbb{R}\ F_n(x) \xrightarrow{\text{a.s.}} F(x)$. The limiting sample spectral c.d.f. F is uniquely determined by c and H; therefore, we will refer to it as $F_{c,H} \coloneqq F$ whenever clarification is needed.

Assumptions 1–3 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 $\nu \geq 1$ of compact intervals: $\mathsf{Supp}(F) = \bigcup_{k=1}^{\nu} [a_k, b_k]$, where $0 < a_1 < b_1 < \cdots < a_\nu < b_\nu < \infty$.

3.3 Hilbert Transform

At this juncture, it is necessary to introduce an important mathematical tool called the *Hilbert transform*. It is defined as convolution with the *Cauchy kernel* $\frac{dt}{t-x}$.

Definition 2. *The Hilbert transform of a real function g is defined as*

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

PV represents the Cauchy Principal Value*, which is used to evaluate the singular integral in the following way:*

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

Recourse to the Cauchy Principal Value is needed because the Cauchy kernel is singular, so the integral does not converge in the usual sense.

Remark 3. Various authors adopt different conventions to define the Hilbert transform. Sometimes the factor $1/\pi$ in front is omitted, and sometimes the kernel is $\frac{dt}{x-t}$. Ultimately, this does not make any difference to the underlying mathematics. Here we adopt the same definition as the monumental recension of known integral transforms published by Erdélyi (1954), based on manuscript notes left by the late CalTech professor Harry Bateman. ■

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 (towards them), but very negative if they are slightly smaller, so it pushes you down (*also* towards them). When the mass points lie far away, it fades out to zero like gravitational attraction does. These effects can be deduced simply by looking at the Cauchy kernel $\frac{dt}{t-x}$. Figure 1 confirms them visually by plotting the Hilbert transform of four well-known densities.

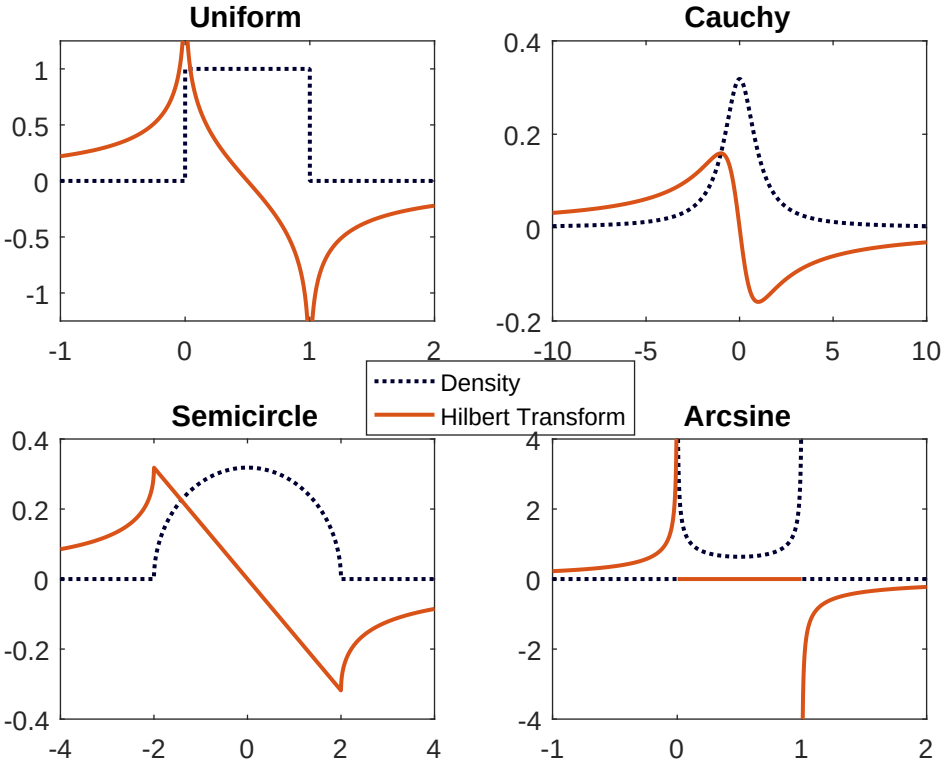

Figure 1: Hilbert Transform of Four Densities. It is strongly positive to the left of the center of mass, strongly negative to the right, and vanishes away from the center of mass.

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. Formulas used in the creation of Figure 1 come from Erdélyi (1954, Chapter XV). They are reproduced for convenience in Table 1.

	Density	Hilbert Transform
Uniform	$f(x) = \mathbb{1}_{\{0 \le x < 1\}}$	$\mathcal{H}_f(x) = \frac{1}{\pi} \log \left\lvert \frac{1-x}{x} \right\rvert$
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	$f(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{\pi\sqrt{x(1-x)}} & x \in (0,1) \\ 0 & x > 1 \end{cases}$	$H_f(x) = \begin{cases} \frac{1}{\pi\sqrt{x(x-1)}} & x < 0 \\ 0 & x \in (0,1) \\ -\frac{1}{\pi\sqrt{x(x-1)}} & x > 1 \end{cases}$

Table 1: Formulas for various densities and their Hilbert transforms.

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 key ingredients in computing the optimal nonlinear shrinkage formula.

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 (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:* $\widehat{\Delta}_n := \mathsf{Diag}\big(\widehat{\delta}_n(\lambda_{n,1}) \ldots, \widehat{\delta}_n(\lambda_{n,p})\big)$*, 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 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 for all* $x \in \mathsf{Supp}(F)$

$$d^{\text{o}}(x) := \frac{x}{\big[\pi c x f(x)\big]^2 + \big[1 - c - \pi c x \mathcal{H}_f(x)\big]^2} \,. \tag{3.2}$$

If Assumptions 1–4 are satisfied, then the following statements hold true:

(a) The oracle estimator of the covariance matrix

$$S_n^{\text{o}} := U_n D_n^{\text{o}} U_n' \quad \textit{where} \quad D_n^{\text{o}} := \mathsf{Diag}\big(d^{\text{o}}(\lambda_{n,1}), \ldots, d^{\text{o}}(\lambda_{n,p})\big) \tag{3.3}$$

minimizes in the class of nonlinear shrinkage estimators defined in Assumption 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 1.*

(b) Conversely, any covariance matrix estimator $\widehat{\Sigma}_n$ *that minimize the a.s. limit of the portfolio selection 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* $\widehat{\delta} = \alpha\, d^{\text{o}}$ *for some positive constant* α*.*

Proof of Theorem 3.1. The results are a direct consequence of Ledoit and Wolf (2017a, Theorem 2) and Engle et al. (2017, Proposition 4.1). ■

3.5 Shrinkage as Local Attraction via the Hilbert Transform

Equation (3.2) was first discovered by Ledoit and Péché (2011, Theorem 3). It 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, Equation (3.2) becomes

$$\forall x \in \mathsf{Supp}(F) \qquad d^o(x) = \frac{x}{1 + c^2\left[\varphi(x)^2 + H_\varphi(x)^2\right] - 2cH_\varphi(x)} \; .$$

This is much more interpretable. If the limiting concentration ratio c is negligible, then the denominator goes to 1, which means no shrinkage. Indeed this is why the sample covariance matrix works well under traditional (fixed-dimensional) asymptotics. As c increases, however, shrinkage must occur. Let us set aside the term $c^2\left[\varphi(x)^2 + H_\varphi(x)^2\right]$ 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 towards 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 towards those of its neighbors that exert the greatest pull. Thus, it could be that it is optimal to "nonlinearly shrink" a relatively small eigenvalue (that is, one that is below average) downwards, if there is a sufficiently massive cluster of slightly inferior eigenvalues attracting it towards them. This effect could never have been anticipated by linear shrinkage. Figure 2 provides a graphical illustration.

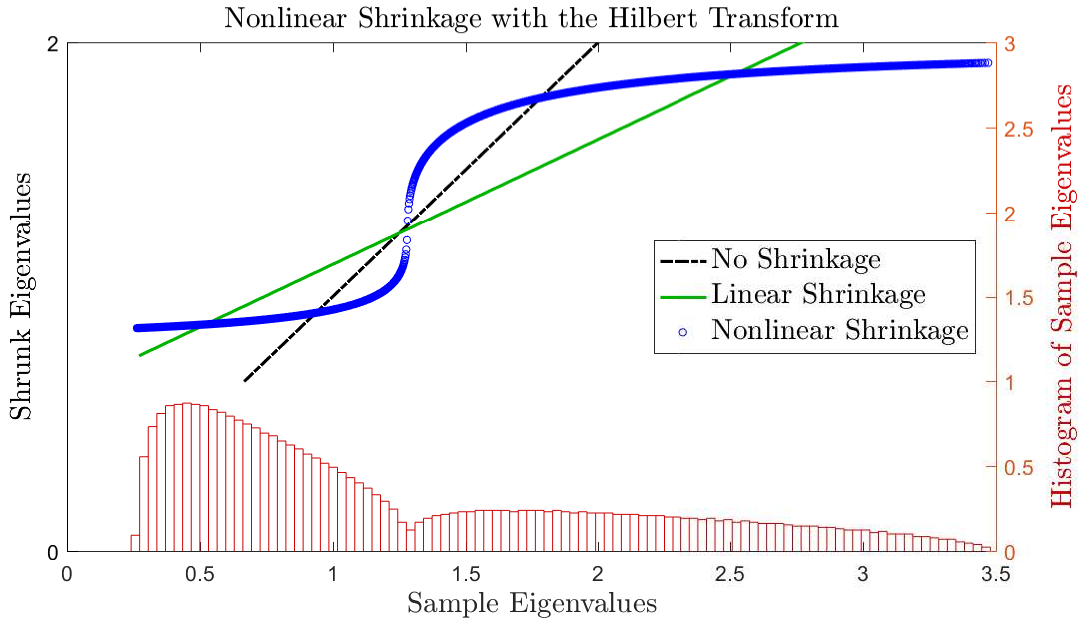

Figure 2: Local Attraction Effect. $2,500$ population eigenvalues are equal to 0.8, and $1,500$ are equal to 2. The sample size is $n = 18,000$. At the bottom of the figure is a histogram displaying the location of the sample eigenvalues.

In this example, the average eigenvalue is equal to 1.25. Sample eigenvalues below the average but above 1 need to be "shrunk" downwards 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. By contrast, nonlinear shrinkage, thanks to its usage of the Hilbert transform, detects local attraction patterns that deviate from the average and adapts accordingly. This is why it is capable of delivering further enhancements over and above those of linear shrinkage.

3.6 Practical Considerations

$\boldsymbol{d}_n^{\text{o}} := \left(d^{\text{o}}(\lambda_{n,1}), \ldots, d^{\text{o}}(\lambda_{n,p})\right)$ represent the large-dimensional counterparts of the finite-sample optimal eigenvalues $\boldsymbol{d}_n^* = \left(d_{n,1}^*, \ldots, d_{n,p}^*\right)$ of Equation (2.2). $\boldsymbol{d}_n^{\text{o}}$ 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 towards the ultimate objective, which is the construction of a *bona fide* estimator (that is, one that can be used in practice) with the same asymptotic properties.

Remark 4. Section 4.2 of Ledoit and Wolf (2017c) proves that the same estimator S_n^{o}

is also optimal within the class of rotation-equivariant estimators of Assumption 4 with respect to the Frobenius loss function. ■

There is considerable interest in estimating the nonlinearly shrunk eigenvalues $\boldsymbol{d}_n^{\text{o}}$ 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 (3.2). 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:

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

Unfortunately, this 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 has been a major stumbling block in the literature. It is the reason why we turn to kernel estimation for the estimation of f and $\mathcal{H}_f$.

4 Asymptotic Theory

4.1 Kernel Requirements

Assumption 5 (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 $i = 1, \ldots, p$, where h_n is a vanishing sequence of positive numbers to be specified below.

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 $\alpha > 0$, then the estimator should remain exactly the same except for rescaling by the same coefficient α. A fixed bandwidth that depends only on n but not on the scale of the eigenvalues would violate this 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.

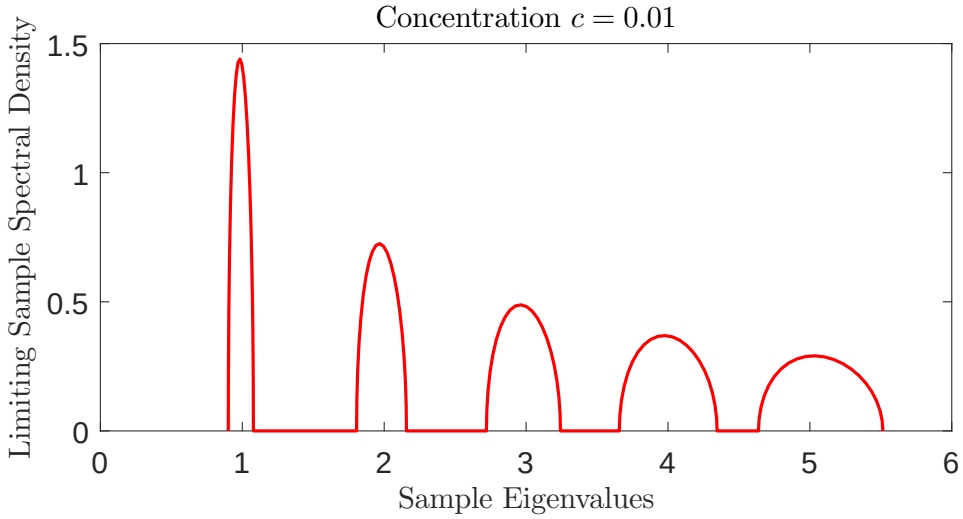

Figure 3: Limiting sample spectral density f when the population eigenvalues are $\{1, 2, \ldots, 5\}$, each with weight $1/5$.

In Figure 3, 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 each 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 estimating the the spectral density.

Additional justification for proportional bandwidth is given by the "arrow model" of Ledoit and Wolf (2017c). 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 error on the sample variance is proportional to the population variance.

4.3 Kernel Estimators

The kernel estimator of the sample spectral p.d.f. f is

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

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

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

4.4 Uniform Consistency

Our main results are as follows. Proofs are in Appendix A.

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*

$$\lim_{n \to \infty} n h_n^{5/2} = \infty \qquad \text{and} \qquad \lim_{n \to \infty} h_n = 0 . \tag{4.1}$$

Moreover, suppose that Assumptions 1–3 are satisfied. Then,

$$\widetilde{f}_n(x) \longrightarrow f(x) \quad \text{and} \quad \mathcal{H}_{\widetilde{f}_n}(x) \longrightarrow \mathcal{H}_f(x) \tag{4.2}$$

in probability uniformly in $x \in \mathsf{Supp}(F)$.

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

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

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

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

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 1–4, the covariance matrix estimator $\widetilde{S}_n$ minimizes in the class of nonlinear shrinkage estimators defined in Assumption 4 the limit in probability of the minimum variance loss function $\mathcal{L}_n^{MV}$, as p and n go to infinity together.*

Remark 5. The above statement remains true if we replace the minimum variance loss function $\mathcal{L}_n^{\text{MV}}$ with the Frobenius loss function $\mathcal{L}_n^{\text{FR}}$.

4.5 Isotonization

Although the $\widetilde{d}_{n,i}$'s have the merit of existing and of being computable from the observable $\boldsymbol{\lambda}_n$, there is no guarantee that they maintain ascending order in finite samples. This is why, in a second phase, we apply the so-called *Pool Adjacent Violators* (PAV) algorithm in order to restore the ascending order. This is the most widespread method of istononization or monotonic regression. It has the advantages of being simple and well-understood, having good theoretical properties, and being easily implementable numerically, with standard code available for free in many programming languages.

The PAV algorithm of Ayer et al. (1955) generates the solution to the equal-weighted isotonic regression problem:

$$\forall \boldsymbol{y} = (y_1, \ldots, y_p) \in \mathbb{R}^p \qquad \mathrm{PAV}(\boldsymbol{y}) := \underset{x_1 \leq \ldots \leq x_p}{\operatorname{argmin}} \frac{1}{p} \sum_{i=1}^{p} (y_i - x_i)^2 \ . \tag{4.4}$$

The isotonization algorithm of Stein (1986) is similar in spirit, but highly non-standard, perhaps due to the fact that it had to handle negative eigenvalues, which we do not, as Equation (3.2) makes clear. From (4.4) we define the estimator

$$\widehat{\boldsymbol{d}}_n = \left(\widehat{d}_{n,1}, \ldots, \widehat{d}_{n,p}\right) := \mathrm{PAV}\left(\widetilde{\boldsymbol{d}}_n\right) \ , \tag{4.5}$$

where $\widetilde{\boldsymbol{d}}_n$ is specified in Section 4.4. Finally, the eigenvalues $\widehat{\boldsymbol{d}}_n$ are stacked into the diagonal of the diagonal matrix $\widehat{D}_n$ to generate the covariance matrix estimator

$$\widehat{S}_n := U_n \widehat{D}_n U_n' = \sum_{i=1}^{p} \widehat{d}_{n,i} \cdot u_{n,i} u_{n,i}' \ . \tag{4.6}$$

This is the estimator that we recommend using in practice because it shrinks the sample eigenvalues in a way that guarantees order preservation.

5 Kernel and Bandwidth Specifications

To operationalize Theorem 4.1, we recommend using the semicircle kernel. It has already been shown by Wigner (1955) to be useful for the eigenvalues of large-dimensional random matrices. Of the 48 elementary functions (that is, not involving the hypergeometric function and other higher transcendental functions) for which the Hilbert transform is known in closed form (Erdélyi, 1954, Section 15.2), the semicircular density is the only one that satisfies Assumption 5.

5.1 Wigner's Semicircle Law

This is the version of the semicircle law that has mean zero and standard deviation one. The most convenient characterization of the semicircle law is through its p.d.f.

$$\forall x \in \mathbb{R} \qquad \kappa(x) = \frac{\sqrt{[4 - x^2]^+}}{2\pi} \ , \tag{5.1}$$

where we have adopted the notation $[y]^+ := \max\{y, 0\}$ for any real y. The support of the p.d.f. is $[-2, 2]$. As mentioned in Table 1, its Hilbert transform is

$$\forall x \in \mathbb{R} \qquad \mathcal{H}_\kappa(x) = \frac{\operatorname{sgn}(x)\sqrt{[x^2 - 4]^+} - x}{2\pi} \ ,$$

where sgn denotes the signum function. A graphical illustration is in the bottom left corner of Figure 1.

Proposition 5.1. *The semicircular kernel satisfies Assumption 5.*

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

$$\widetilde{f}_n(\lambda_{n,i}) = \frac{1}{p} \sum_{j=1}^{p} \frac{\sqrt{\left[4\lambda_{n,j}^2 h_n^2 - (\lambda_{n,i} - \lambda_{n,j})^2\right]^+}}{2\pi\lambda_{n,j}^2 h_n^2} \tag{5.2}$$

$$\mathcal{H}_{\widetilde{f}_n}(\lambda_{n,i}) = \frac{1}{p} \sum_{j=1}^{p} \frac{\operatorname{sgn}(\lambda_{n,i} - \lambda_{n,j})\sqrt{\left[(\lambda_{n,i} - \lambda_{n,j})^2 - 4\lambda_{n,j}^2 h_n^2\right]^+} - \lambda_{n,i} + \lambda_{n,j}}{2\pi\lambda_{n,j}^2 h_n^2} \ . \tag{5.3}$$

5.2 Choice of Bandwidth

There are two possible sources of inspiration for the speed at which the bandwidth vanishes in n. The first is the standard kernel density estimation theory that recommends a bandwidth of order $n^{-1/5}$; for example, see Silverman (1986).

The second is the so-called *Arrow Model* of Ledoit and Wolf (2017c). In Lemmas E.4 and E.7, they show that, if there is a single isolated population eigenvalue τ_p of large order of magnitude, the width of the interval where the corresponding sample eigenvalue can lie is of the order $4\tau_p/\sqrt{n}$. This points instead to a bandwidth control parameter of order $n^{-1/2}$.

In order to strike a compromise between these two disparate approaches, we take the average of the two exponents:

$$\frac{\frac{1}{5} + \frac{1}{2}}{2} = 0.35 \ .$$

Further justification for the choice of 0.35 as exponent comes from Jing et al. (2010, Theorem 1): Only exponents strictly below 0.40 guarantee convergence of the kernel

density estimator. Indeed, 0.35 is very close to the exponent of 1/3 chosen by Jing et al. (2010) themselves in the simulation study of their Section 4. Thus, we set henceforth

$$h_n := n^{-0.35} \ . \tag{5.4}$$

As mentioned in Section 4.2, we need $h_n < 1/R$, where R is half the width of the support of the kernel — in this case, 2. With (5.4) this is achieved as soon as $n \geq 8$. Given the asymptotic nature of our results, we will henceforth implicitly assume that $n \geq 8$ in all our calculations.

6 Monte Carlo Simulations

6.1 Competitors

We compare the performance of six covariance matrix estimators:

Sample The sample covariance matrix S_n.

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

Direct The direct kernel estimator $\widehat{S}_n$ of Equation (4.6).

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 splitting the data.

FSOPT The finite-sample optimal estimator Σ_n^* defined in Equation (2.3), which would require knowledge of the unobservable population covariance matrix Σ_n, and thus is not applicable in the real world, but is a useful benchmark.

The Matlab code for NERCOME was generously provided by Professor Clifford Lam from the Department of Statistics at the London School of Economics. We are also grateful to Dr. Sean R. Collins for having made publicly available the Matlab function `pav.m` that implements the PAV algorithm as part of his E-MAP toolbox (Collins et al., 2006). The code for the QuEST package comes from the numerical implementation detailed in Ledoit and Wolf (2017b).

6.2 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

$$\text{PRIAL}_n^{\text{MV}}\big(\widehat{\Sigma}_n\big) \coloneqq \frac{\mathbb{E}\big[\mathcal{L}_n^{\text{MV}}\big(S_n, \Sigma_n\big)\big] - \mathbb{E}\big[\mathcal{L}_n^{\text{MV}}\big(\widehat{\Sigma}_n, \Sigma_n\big)\big]}{\mathbb{E}\big[\mathcal{L}_n^{\text{MV}}\big(S_n, \Sigma_n\big)\big] - \mathbb{E}\big[\mathcal{L}_n^{\text{MV}}\big(S_n^*, \Sigma_n\big)\big]} \times 100\% \;, \tag{6.1}$$

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 by Proposition 2.1. Given that the construction of S_n^* requires knowledge of the unobservable population covariance matrix Σ_n, 100% improvement represents an upper limit that is unattainable in reality. The question is how close to the speed-of-light of 100% a *bona fide* estimator can get.

Recall that the loss function $\mathcal{L}_n^{\text{MV}}$ 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. Higher PRIAL is better, obviously.

6.3 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 4.00GHz Windows desktop PC running Matlab R2016a.

Estimator	Sample	Linear	Direct	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	3	4	2,233	2,990	3

Table 2: Simulation results for the baseline scenario.

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/Direct/QuEST), which is a very satisfactory number.

One key lesson is that the direct kernel approach championed in the present paper is faster than all the other nonlinear shrinkage methods by two orders of magnitude. Thus, direct kernel shrinkage delivers the best of both worlds: QuEST-tier variance reduction at Linear-tier speed. Note also that 3 of the 4 milliseconds spent on direct estimation are spent on extracting the eigenvalues and eigenvectors of the sample covariance matrix, an operation that all nonlinear shrinkage methods must perform, even if they know the true covariance matrix (cf. FSOPT).

The only estimator that is in the same ballpark as direct kernel shrinkage in terms of both speed and accuracy is the finite-sample optimal estimator, which presupposes foreknowledge of the true covariance matrix, an unrealistic assumption. Among *bona fide* estimators, the direct kernel estimator is the only one that comes even close to matching both the speed and accuracy of the finite-sample optimal estimator.

Table 2 shows that statisticians who are already comfortable with linear shrinkage and would like to upgrade to nonlinear for performance enhancement, but have been concerned by the numerical complexity of the earlier techniques, can now safely switch to direct kernel estimation.

6.4 Convergence

6.4.1 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.

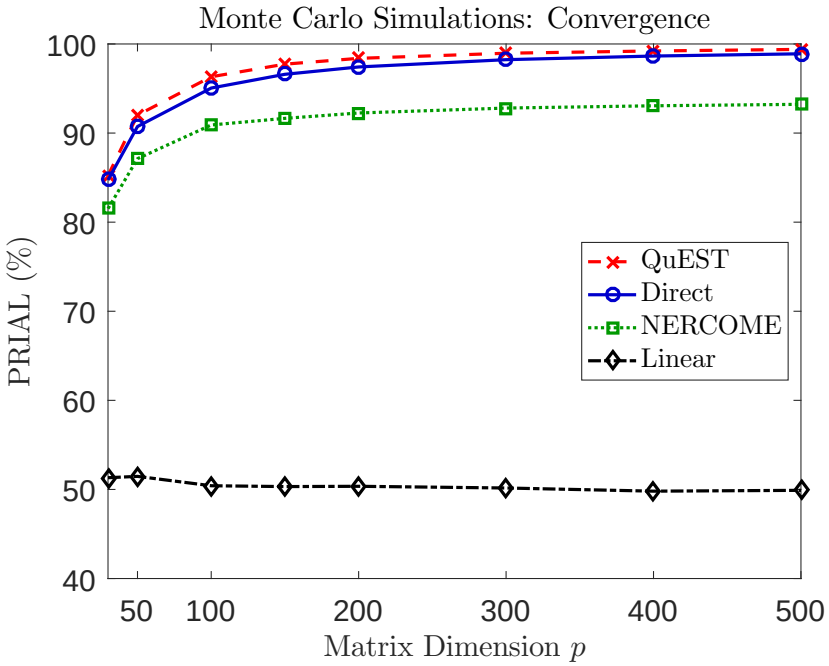

Figure 4: Evolution of the PRIAL of various estimators as the matrix dimension and the sample size go to infinity together.

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

6.4.2 Speed

Apart from minimizing the expected loss, a key advantage of the direct kernel 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.

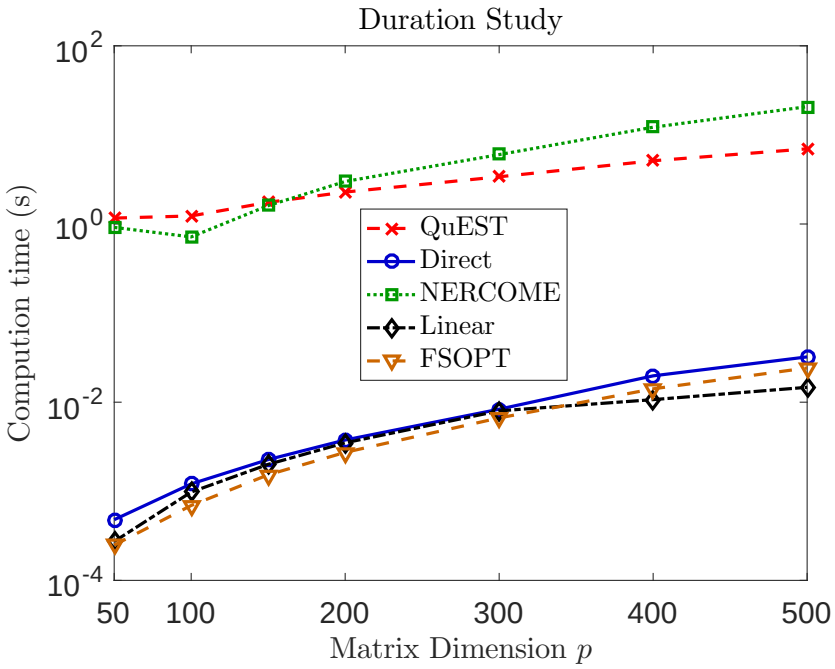

Figure 5: Computational speed of various shrinkage estimators as the matrix dimension and the sample size go to infinity together, measured in seconds, with log-scale on the vertical axis.

There is a clear gap between, on the one hand, QuEST and NERCOME and, on the other hand, Direct, Linear and Oracle. The Direct estimator is faster than the other nonlinear shrinkage estimators by a factor of more than 200.

6.4.3 Ultra-High Dimension

The direct kernel estimation method 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.

Estimator	Sample	Linear	Direct	FSOPT
Average Loss	2.679	2.086	1.488	1.487
PRIAL	0%	49.74%	99.92%	100%
Time (s)	21	43	113	108

Table 3: Result of 100 Monte Carlo simulations for dimension $p = 10,000$ and sample size $n = 30,000$.

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

Speed-wise, it takes less than two minutes to compute direct nonlinear shrinkage 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 (4.6) (also $O(p^3)$ cost). These operations would be necessary for any nonlinear shrinkage estimator — even if we knew the unobservable 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 $\mathcal{H}_f$ as defined in Section 4.3 and of the shrunk eigenvalues themselves (4.5), 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 simulations (not reported here) in dimension $p = 20,000$ with sample size $n = 60,000$ show computation times 7.6 to 8.1 times longer for the four estimators of Table 3, which tightly brackets the theoretical prediction of $2^3 = 8$.

6.5 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.

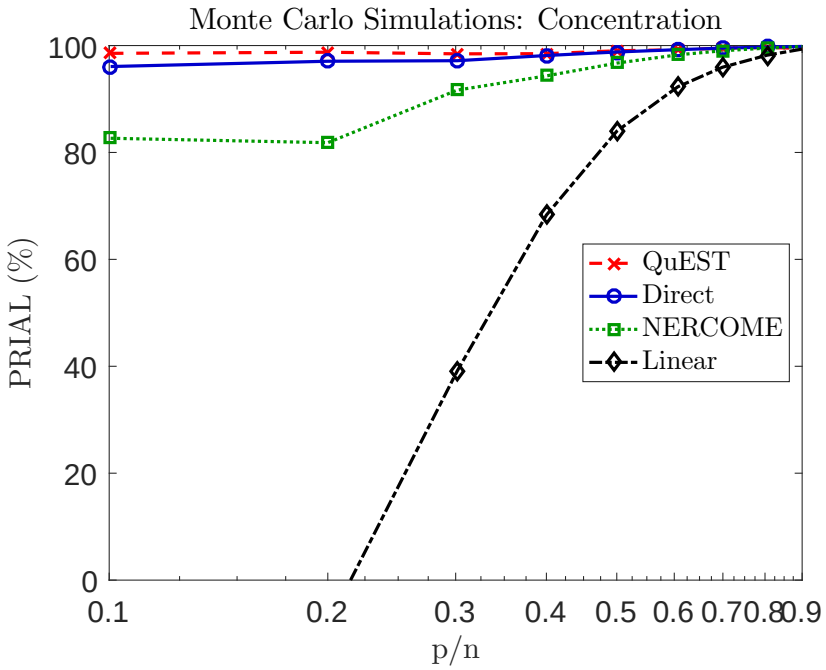

Figure 6: Evolution of the PRIAL of various estimators as a function of the ratio of the matrix dimension to the sample size.

Linear shrinkage performs very well in high concentrations, but does not beat the sample covariance matrix for low concentrations. Appendix B.1 shows that this is solely 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 would always beats the sample covariance matrix in the same simulation experiment.

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

6.6 Condition Number

We start again from the baseline 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.

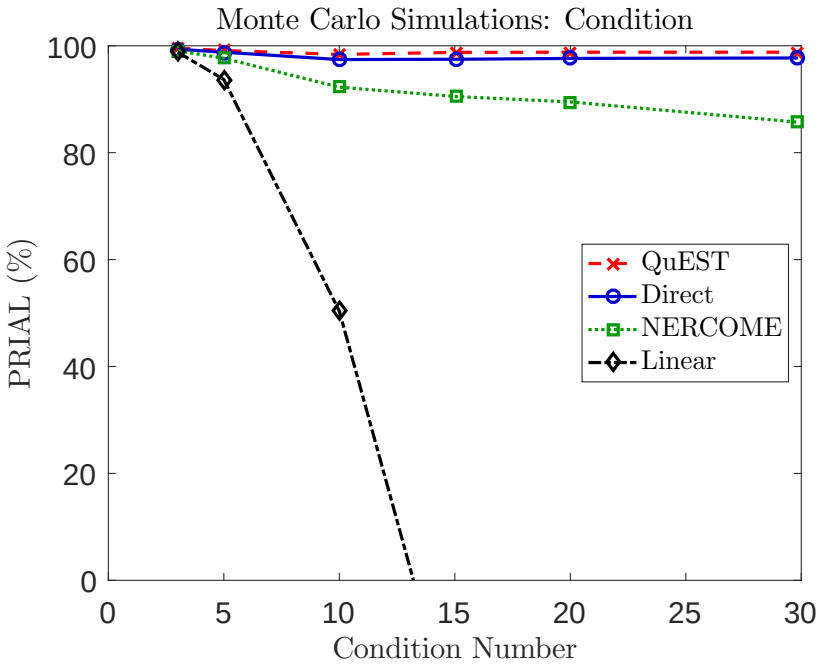

Figure 7: Evolution of the PRIAL of various estimators as a function of the condition number of the population covariance matrix.

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 verifies it 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 formulas all capture a very high percentage of the potential for variance reduction, with Direct in particular being very close to QuEST and above the 97% mark across the board.

6.7 Non-Normality

In this experiment, we start from the baseline scenario and change the distribution of the variates. We study the Bernoulli coin toss distribution, which is the most platykurtic of all distributions, the Laplace distribution, which is leptokurtotic, and the "Student" t-distribution with 5 degrees of freedom, also leptokurtotic. All of these are suitably normalized to have mean zero and variance one, if necessary. The results are presented in Table 4.

Distribution	Linear	Direct	QuEST	NERCOME
Bernoulli	51%	98%	98%	92%
Laplace	50%	97%	98%	92%
'Student' t_5	49%	97%	98%	92%

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

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

6.8 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.

Beta Parameters	Linear	Direct	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%

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

Note that the 100% PRIALs are solely 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 formulas capture a very high percentage of the potential for variance reduction in all cases, with Direct being virtually indistinguishable from QuEST and above the 97% mark across the board.

6.9 Fixed-Dimension Asymptotics

An instructive experiment that falls outside the purview of large-dimensional asymptotics is to keep the dimension p constant 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.

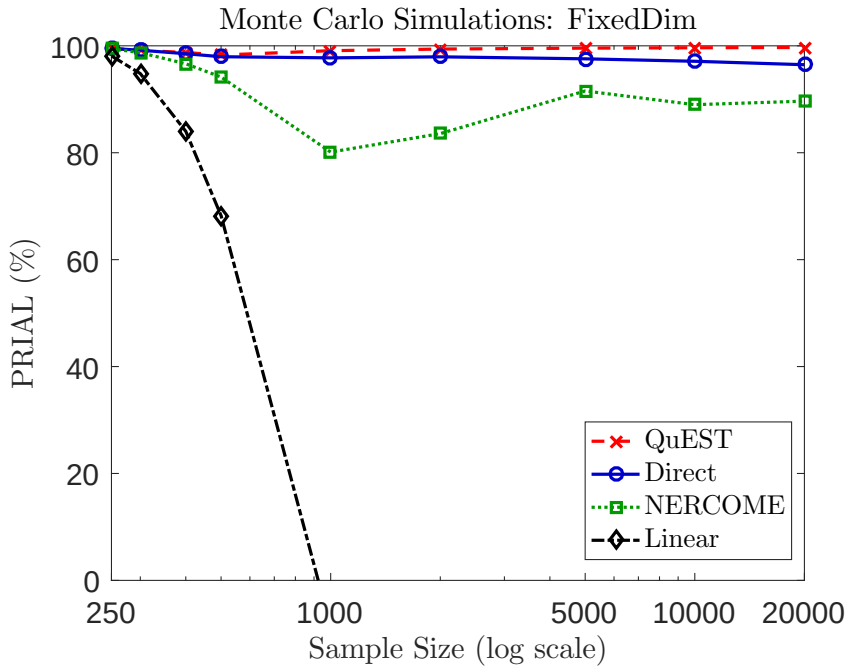

Figure 8: Evolution of the PRIAL as the sample size grows towards infinity, while the matrix dimension remains fixed.

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) concentration ratios. Appendix B.3 shows that linear shrinkage does not suffer from any such weakness in the Frobenius loss.

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

6.10 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 (2017c, Section 7), we set the largest eigenvalue (the 'arrow' eigenvalue) 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 $[1, 10]$. The results are displayed in Figure 9, where the matrix dimension varies from $p = 50$

to $p = 500$.

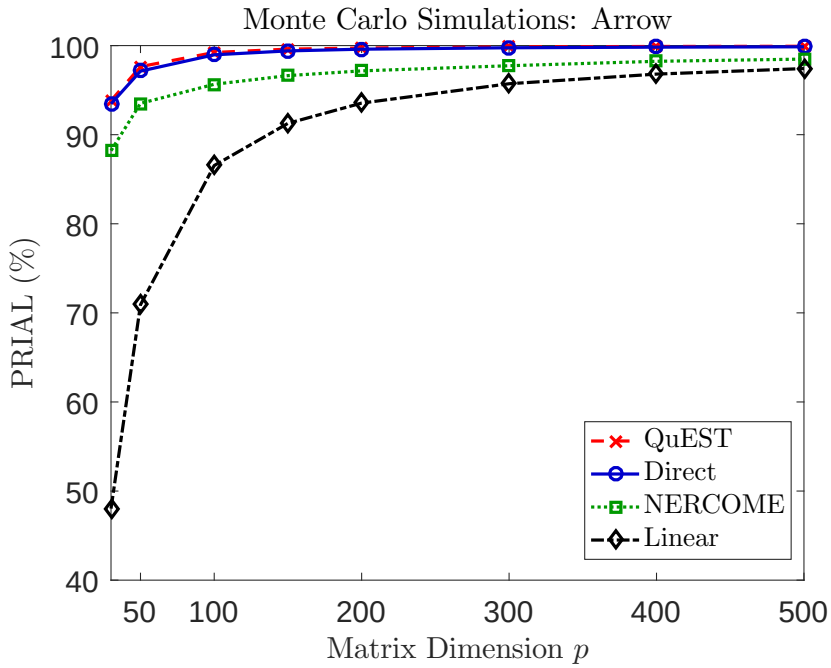

Figure 9: Evolution of the PRIAL as the sample size grows towards infinity, while the matrix dimension remains fixed.

All shrinkage formulas seem to be able to handle the arrow eigenvalue of order p competently. The three nonlinear shrinkage estimators have the highest performance, and Direct is always above the 97% mark.

6.11 Variants

Additional simulations (not reported here) show that post-processing the shrunk eigenvalues $\widetilde{\boldsymbol{d}}_n$ via the PAV algorithm generates very little improvement. For example, in the baseline scenario, the PRIAL of the direct kernel estimator is 98.16% without PAV and 98.24% with PAV, which is a difference of 0.08%, negligible in practice. This scenario is actually representative of the order of magnitude of the difference across the board. The difference is always positive but never exceeds 0.77%. The charts with and without PAV are indistinguishible to the naked eye. Thus, PAV is not an essential part of our estimator, it just falls under "nice to have". This verification is important because isotonization *is* an essential part of Stein's (1986) nonlinear shrinkage estimator, which is problematic because its mathematical properties are so hard to investigate; for example, see Rajaratnam and Vincenzi (2016).

By contrast, making the bandwidth proportional to the sample eigenvalues, as we do in Section 4.2, as opposed to making it independent from them (or uniform) is crucial. For example, in the baseline scenario, the PRIAL of the direct kernel estimator is 98%

with proportional bandwidth and 80% with uniform bandwidth[1], which is a difference of 18%, substantial in practice. This scenario is actually representative of the order of magnitude of the difference across the board. The difference is always positive and can even reach 55% when the condition number is high. Thus, the mathematical work that extends the kernel estimator of Jing et al. (2010) from uniform to proportional bandwidth is an essential part of our covariance matrix estimator.

6.12 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 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 Direct, which tends to be better than NERCOME, but the differences are relatively small, and there are exceptions. Between QuEST and Direct there is even hardly any difference.

The direct kernel estimator is very simple to implement, as proven by the 20-line Matlab code in Appendix D. It captures 90% or more of the potential for variance reduction that comes from shrinking the sample eigenvalues. It is 200 times faster than the other nonlinear shrinkage methods, and is the only one that can handle ultra-high dimensions up to $10,000$ in reasonable time.

7 Conclusion

The contribution of this paper has been to make the power of nonlinear shrinkage as easily accessible as that of linear shrinkage. We have achieved this by greatly simplifying the estimation technology. The key 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. The resulting computations are easy to understand, straightforward to implement, fast, and scalable.

Extensive Monte Carlo simulations show that the direct kernel 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 so-called "curse of dimensionality" which is often

[1]The uniform-bandwidth method uses $h_{n,i} \equiv \bar{\lambda}_n n^{-0.35}$, where $\bar{\lambda}_n$ is the grand mean of the $\lambda_{n,i}$.

associated with portfolio selection involving large covariance matrices of stock returns.

The dimension of covariance matrices that can be handled successfully now is one order of magnitude larger compared to earlier nonlinear shrinkage methods, which is important in the age of big data. We trust that this feature will make nonlinear shrinkage even more attractive to applied researchers.

A Proofs

The way to prove Theorem 4.1 is to extend the proof of Theorem 1 in Jing et al. (2010). As a result, the first priority is to shift from the Hilbert transform, which is the mathematical tool favored in the main body of the text of our own paper, to a closely related complex transform called the Stieltjes (1894) transform, the instrument utilized by Jing et al. (2010). Like the Hilbert transform, the Stieltjes transform convolves with the Cauchy kernel. The only difference is that its argument lies in $\mathbb{C}^+$, the half-plane of complex numbers with strictly positive imaginary part, whereas the argument of the Hilbert transform is instead a real number.

A.1 Stieltjes Transform

Given any c.d.f. G, its Stieljes transform m_G is defined as

$$\forall z \in \mathbb{C}^+ \qquad m_G(z) \coloneqq \int_{-\infty}^{+\infty} \frac{1}{x - z} dG(x) \ .$$

When G is sufficiently regular, its Stieltjes transform admits an extension to the real line, which we denote as

$$\breve{m}_G(x) \coloneqq \lim_{z \in \mathbb{C}^+ \to x} m_G(z) \qquad \text{for all } x \in \mathbb{R} \ .$$

Note that, although $\breve{m}_G$ is a function of real argument, it is generally complex-valued. Both its real and imaginary parts have nice interpretations, as the following equation shows:

$$\breve{m}_G(x) = \pi \left[\mathcal{H}_{G'}(x) + \sqrt{-1}\, G'(x) \right] \ .$$

Thus, any statement about the extension to the real line of the Stieltjes transform of a c.d.f. is really a statement about the corresponding p.d.f. and its Hilbert transform.

Under Assumptions 1–3, Theorem 1.1 of Silverstein and Choi (1995) implies that $\breve{m}_F(x)$ exists and is continuous. Our approach is to estimate it with the kernel estimator

$$\forall x \in \mathbb{R} \qquad \breve{m}_{\widetilde{F}_n}(x) \coloneqq \frac{1}{p} \sum_{i=1}^{p} \frac{1}{h_{n,i}} \breve{m}_K \left(\frac{x - \lambda_{n,i}}{h_{n,i}} \right) = \lim_{z \in \mathbb{C}^+ \to x} \int \frac{\widetilde{f}_n(t)}{t - z} dt \ ,$$

where $\widetilde{F}_n$ is the kernel estimator of the limiting sample spectral c.d.f. defined by

$$\forall x \in \mathbb{R} \qquad \widetilde{F}_n(x) \coloneqq \frac{1}{p} \sum_{i=1}^{p} K \left(\frac{x - \lambda_{n,i}}{h_{n,i}} \right) = \int_{-\infty}^{x} \widetilde{f}_n(t) dt \ ,$$

and K is the kernel's c.d.f.: $K(x) \coloneqq \int_{-\infty}^{x} k(t) dt$.

A.2 Assumptions

The assumptions in our paper are couched in slightly different terms than those made by Jing et al. (2010). Therefore it is necessary, before proceeding any further, to establish that one set of assumptions maps into the other.

First, given that Assumption 5 requires the kernel k to be continuous with compact support, Jing et al.'s (2010) Equation (2.3)

$$\sup_{-\infty<x<\infty} |k(x)| < \infty, \qquad \lim_{|x|\to\infty} |xk(x)| = 0$$

is satisfied. In addition, their Equation (2.4)

$$\int k(x)dx = 1, \qquad \int |k'(x)|dx < \infty$$

is satisfied because Assumption 5 requires k to be a p.d.f. on the one hand, and a function of bounded variation on the other hand. Therefore the assumptions of Jing et al.'s (2010) Theorem 1 are satisfied here.

In what follows, we will require a slightly stronger assumption, namely:

$$\int \left| \frac{d\breve{m}_K}{dx}(x) \right| dx < \infty \ . \tag{A.1}$$

The imaginary part has already been taken care of, as it is π times the kernel density. As for the real part, it follows from the statement in Assumption 5 that requires the Hilbert transform $\mathcal{H}_k$ to be a function of bounded variation. So (A.1) holds. From now on, we define $\breve{m}'_K(x) := \frac{d\breve{m}_K}{dx}(x)$.

A.3 Lemmas

Lemma A.1. *Under the assumptions of Theorem 4.1, let $F_{c_n,H_n}(x)$ be the c.d.f. obtained from $F_{c,H}(x)$ by replacing c and H with c_n and H_n, respectively. Furthermore, let $\breve{m}_{F_{c_n,H_n}}(x)$ denote the extension to the real line of the Stieltjes transform of $F_{c_n,H_n}(x)$. Then,*

$$\sup_{n,x} \left| \breve{m}_{F_{c_n,H_n}}(x) \right| < \infty \ . \tag{A.2}$$

Proof of Lemma A.1. Lemma A.1 follows immediately from Equation (5.5) of Jing et al. (2010). ■

Lemma A.2. $\lim_{x\to+\infty} \frac{1}{x} \int_{-x}^{x} |\breve{m}_K(t)|dt = 0$.

Proof of Lemma A.2.

$$\forall x \geq R+1 \qquad \int_{R+1}^{x} |\breve{m}_K(t)|dt = \int_{R+1}^{x} \int_{-R}^{R} \frac{k(u)}{t-u} du\, dt \leq \int_{R+1}^{x} \frac{1}{t-R} dt = \log(x-R) \ ,$$

therefore

$$\frac{1}{x}\int_{-x}^{x}|\breve{m}_K(t)|dt \leq \frac{1}{x}\left[2\log(x-R)+\int_{-R-1}^{R+1}|\breve{m}_K(t)|dt\right] ,$$

which vanishes as $x \to +\infty$. ■

A.4 Proof of Theorem 4.1

First, we claim that

$$\sup_{x\in[a,b]}\left|\breve{m}_{\widetilde{F}_n}(x)-\int_a^b \frac{1}{th_n}\breve{m}_K\left(\frac{x-t}{th_n}\right)dF_{c_n,H_n}(t)\right| \longrightarrow 0 \tag{A.3}$$

in probability. Indeed, from integration by parts and Theorem 3 of Jing et al. (2010),

$$\begin{aligned}
&\mathbb{E}\sup_{x\in[a,b]}\left|\int_a^b \frac{1}{th_n}\breve{m}_K\left(\frac{x-t}{th_n}\right)dF_n(t)-\int_a^b \frac{1}{th_n}\breve{m}_K\left(\frac{x-t}{th_n}\right)dF_{c_n,H_n}(t)\right| \\
&\quad=\mathbb{E}\sup_{x\in[a,b]}\left|\int_a^b \frac{1}{t^2h_n}\left[\breve{m}_K\left(\frac{x-t}{th_n}\right)+\frac{x}{th_n}\breve{m}'_K\left(\frac{x-t}{th_n}\right)\right]\times\left[F_n(t)-F_{c_n,H_n}(t)\right]dt\right| \\
&\quad=\mathbb{E}\sup_{x\in[a,b]}\left|\int_{\frac{x-b}{bh_n}}^{\frac{x-a}{ah_n}}\frac{(1+uh_n)^2}{x^2h_n}\left[\breve{m}_K(u)+\frac{1+uh_n}{h_n}\breve{m}'_K(u)\right]\right. \\
&\qquad\qquad\left.\times\left[F_n\left(\frac{x}{1+uh_n}\right)-F_{c_n,H_n}\left(\frac{x}{1+uh_n}\right)\right]\frac{xh_n}{(1+uh_n)^2}du\right| \\
&\quad=\mathbb{E}\sup_{x\in[a,b]}\left|\int_{\frac{x-b}{bh_n}}^{\frac{x-a}{ah_n}}\frac{1}{x}\left[\breve{m}_K(u)+\frac{1+uh_n}{h_n}\breve{m}'_K(u)\right]\right. \\
&\qquad\qquad\left.\times\left[F_n\left(\frac{x}{1+uh_n}\right)-F_{c_n,H_n}\left(\frac{x}{1+uh_n}\right)\right]du\right| \\
&\quad\leq\frac{1}{h_n}\mathbb{E}\sup_x|F_n(x)-F_{c_n,H_n}(x)|\times\left[h_n\int_{\frac{a-b}{ah_n}}^{\frac{b-a}{ah_n}}|\breve{m}_K(u)|\,du+\frac{b}{a}\int_{-\infty}^{+\infty}|\breve{m}'_K(u)|\,du\right] \\
&\quad=O\left(\frac{1}{n^{2/5}h_n}\right)\longrightarrow 0 .
\end{aligned}$$

The next aim is to show that

$$\int\frac{1}{th_n}\breve{m}_K\left(\frac{x-t}{th_n}\right)dF_{c_n,H_n}(t)-\int\frac{1}{th_n}\breve{m}_K\left(\frac{x-t}{th_n}\right)dF_{c,H}(t)\longrightarrow 0 \tag{A.4}$$

uniformly in $x \in \mathsf{Supp}(F)$. This is equivalent to, for any sequence $\{x_n, n \geq 1\}$ in $\mathsf{Supp}(F)$ converging to x,

$$\int\frac{1}{1+uh_n}\breve{m}_K(u)\left[F'_{c_n,H_n}\left(\frac{x_n}{1+uh_n}\right)-F'_{c,H}\left(\frac{x_n}{1+uh_n}\right)\right]du\longrightarrow 0 . \tag{A.5}$$

From Theorem 1.1 of Silverstein and Choi (1995), $F'_{c,H}$ is uniformly bounded on $\mathsf{Supp}(F)$. Therefore, (A.5) follows from the dominated convergence theorem, Lemma A.1 and Lemma 2 of Jing et al. (2010).

The final step is divided into two sub-items, by considering the real part (which is the Hilbert transform of the density) and the imaginary part (which is the density itself) separately. Recall that PV denote the Cauchy Principal Value of an improper integral. Regarding the real part, we observe that

$$\begin{aligned}
\int \frac{1}{th_n}\mathsf{Re}\left[\breve{m}_K\left(\frac{x-t}{th_n}\right)\right] dF_{c,H}(t) &= \int \frac{1}{th_n} PV \int_{-R}^{R} \frac{k(v)}{v-\frac{x-t}{th_n}} dF_{c,H}(t) \\
&= \int_{-R}^{R} k(v) PV \int \frac{1}{th_n} \frac{F'_{c,H}(t)}{v-\frac{x-t}{th_n}} dt dv \\
&= \int_{-R}^{R} \frac{1}{1+vh_n} k(v) PV \int \frac{F'_{c,H}(t)}{t-\frac{x}{1+vh_n}} dt dv \\
&= \int_{-R}^{R} \frac{1}{1+vh_n} k(v) \mathsf{Re}\left[\breve{m}_{F_{c,H}}\left(\frac{x}{1+vh_n}\right)\right] dv \ .
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sup_{x\in[a,b]} \left| \int \frac{1}{th_n}\mathsf{Re}\left[\breve{m}_K\left(\frac{x-t}{th_n}\right)\right] dF_{c,H}(t) - \mathsf{Re}[\breve{m}_{F_{c,H}}(x)] \right| \\
&\quad = \sup_{x\in[a,b]} \left| \int_{-R}^{R} \frac{1}{1+vh_n} k(v) \mathsf{Re}\left[\breve{m}_{F_{c,H}}\left(\frac{x}{1+vh_n}\right)\right] dv - \mathsf{Re}[\breve{m}_{F_{c,H}}(x)] \right| \\
&\quad \leq \sup_{x\in[a,b]} \left| \int_{-R}^{R} \frac{1}{1+vh_n} k(v) \left\{ \mathsf{Re}\left[\breve{m}_{F_{c,H}}\left(\frac{x}{1+vh_n}\right)\right] - \mathsf{Re}[\breve{m}_{F_{c,H}}(x)] \right\} dv \right| \\
&\qquad + \sup_{x\in[a,b]} \left| \mathsf{Re}[\breve{m}_F(x)] \right| \times \left| 1 - \int_{-R}^{R} \frac{1}{1+vh_n} k(v) dv \right| \ . \qquad \text{(A.6)}
\end{aligned}$$

Note that, by the dominated convergence theorem,

$$\lim_{n\to\infty} \int_{-R}^{R} \frac{1}{1+vh_n} k(v) dv = \int_{-R}^{R} k(v) dv = 1 \ . \qquad \text{(A.7)}$$

Therefore, the second term in (A.6) is $o(1)$. And for the first term it holds that

$$\begin{aligned}
&\sup_{x\in[a,b]}\left|\int_{-R}^{R}\frac{1}{1+vh_n}k(v)\left\{\mathsf{Re}\left[\breve{m}_{F_{c,H}}\left(\frac{x}{1+vh_n}\right)\right]-\mathsf{Re}[\breve{m}_{F_{c,H}}(x)]\right\}dv\right| \\
&\quad\leq \sup_{x\in[a,b]}\int_{-R}^{R}\frac{1}{1+vh_n}k(v)\left|\mathsf{Re}\left[\breve{m}_{F_{c,H}}\left(\frac{x}{1+vh_n}\right)\right]-\mathsf{Re}[\breve{m}_{F_{c,H}}(x)]\right|dv \\
&\quad\leq \sup_{\substack{x,y\in\left[a-\frac{Rh_n}{1-Rh_n},b+\frac{Rh_n}{1-Rh_n}\right]\\ |x-y|\leq\frac{Rh_n}{1-Rh_n}}}\left|\mathsf{Re}\left[\breve{m}_{F_{c,H}}(y)\right]-\mathsf{Re}[\breve{m}_{F_{c,H}}(x)]\right|\times\left|\int_{-R}^{R}\frac{1}{1+vh_n}k(v)dv\right| \\
&\quad\leq \sup_{\substack{x,y\in\left[\frac{a}{2},b+\frac{a}{2}\right]\\ |x-y|\leq\frac{Rh_n}{1-Rh_n}}}\left|\mathsf{Re}\left[\breve{m}_{F_{c,H}}(y)\right]-\mathsf{Re}[\breve{m}_{F_{c,H}}(x)]\right|\times\left|\int_{-R}^{R}\frac{1}{1+vh_n}k(v)dv\right| \text{ for large } n.
\end{aligned} \tag{A.8}$$

By the Heine-Cantor theorem, the first term of expression (A.8) converges to zero and by (A.7) the second expression of (A.8) converges to one. This guarantees that the bound (A.8) converges to zero as $n\to\infty$. This ends the proof for the real part.

Concerning the proof for the imaginary part, the statement we seek to establish is

$$\sup_{x\in[a,b]}\left|\int\frac{1}{th_n}\mathsf{Im}\left[\breve{m}_K\left(\frac{x-t}{th_n}\right)\right]dF_{c,H}(t)-\mathsf{Im}[\breve{m}_{F_{c,H}}(x)]\right|\longrightarrow 0\ . \tag{A.9}$$

A closely related statement, namely

$$\sup_{x\in[a,b]}\left|\int\frac{1}{h_n}\mathsf{Im}\left[\breve{m}_K\left(\frac{x-t}{h_n}\right)\right]dF_{c,H}(t)-\mathsf{Im}[\breve{m}_{F_{c,H}}(x)]\right|\longrightarrow 0\ , \tag{A.10}$$

was proven by Jing et al. (2010) in the course of proving their Theorem 1 at the end of Section 5.1. It can be verified that their method of proof can be adapted to establish the truth of (A.9), using the techniques developed above for the real part. The adaptation is not mathematically difficult, as all the hard work has been already done by Jing et al. (2010). But the details are tedious, so they are left to the reader.

Note that (A.8) and (A.9) together imply

$$\sup_{x\in[a,b]}\left|\int\frac{1}{th_n}\breve{m}_K\left(\frac{x-t}{th_n}\right)dF_{c,H}(t)-\breve{m}_{F_{c,H}}(x)\right|\longrightarrow 0\ . \tag{A.11}$$

Results (A.3), (A.4), and (A.11) together conclude the proof of Theorem 4.1. ■

A.5 Proof of Corollary 4.1

By Theorem 4.1, the shrinkage function

$$x\longmapsto\frac{x}{\left|1-(p/n)-(p/n)x\breve{m}_{\widetilde{F}_n}(x)\right|^2}$$

converges in probability to the oracle shrinkage function $d^{\mathrm{o}}(x)$ for all $x \in \mathsf{Supp}(F)$. Therefore, the estimator $\widetilde{S}_n$ has the same asymptotic loss as the oracle S_n^{o}, which is the minimum in its class by Theorem 3.1. ■

A.6 Proof of Proposition 5.1

Simply from its definition, it is obvious that the semicircular kernel (5.1) is a p.d.f., is continuous, symmetric, and nonnegative, that it has mean zero, variance one, and that its support is a compact interval. The fact that its Hilbert transform exists and is continuous follows from Equation (19) of Erdélyi (1954, Section 15.2).

The semicircular kernel (5.1) is a function of bounded variation because it is increasing on $[-2, 0]$, decreasing on $[0, 2]$, and constant everywhere else. Similarly, its Hilbert transform is a function of bounded variation because it is increasing on $(-\infty, -2)$, decreasing on $[-2, 2]$, and increasing again on $(2, \infty)$, which implies that

$$\begin{aligned}\int_{-\infty}^{+\infty} \left| \frac{d\mathcal{H}_\kappa}{dx}(x) \right| dx &= \int_{-\infty}^{-2} \frac{d\mathcal{H}_\kappa}{dx}(x)dx - \int_{-2}^{2} \frac{d\mathcal{H}_\kappa}{dx}(x)dx + \int_{2}^{+\infty} \frac{d\mathcal{H}_\kappa}{dx}(x)dx \\ &= -\lim_{x \to -\infty} \mathcal{H}_\kappa(x) + 2\mathcal{H}_\kappa(-2) - 2\mathcal{H}_\kappa(2) + \lim_{x \to +\infty} \mathcal{H}_\kappa(x) \\ &= \frac{4}{\pi} < \infty \ . \ \blacksquare\end{aligned}$$

B Frobenius Loss

The linear shrinkage estimator of Ledoit and Wolf (2004) has two scalar parameters that are optimized with respect to the Frobenius loss. Given the poor performance of linear shrinkage in Sections 6.5, 6.6, and 6.9 — namely, its inability to dominate the sample covariance matrix over certain parts of the parameter space in terms of the Minimum Variance loss function — it is important to verify that this is solely due to the fact that linear shrinkage has been unfairly handicapped by the switch of loss function. The many applied statisticians who use linear shrinkage because they believe that it improves over the sample covariance matrix need to be reassured about its performance. The key quantity in this investigation is the Frobenius PRIAL We define the Frobenius PRIAL in manner analogous to Equation (6.1) as

$$\mathrm{PRIAL}_n^{\mathrm{FR}}\big(\widehat{\Sigma}_n\big) := \frac{\mathbb{E}\big[\mathcal{L}_n^{\mathrm{FR}}\big(S_n, \Sigma_n\big)\big] - \mathbb{E}\big[\mathcal{L}_n^{\mathrm{FR}}\big(\widehat{\Sigma}_n, \Sigma_n\big)\big]}{\mathbb{E}\big[\mathcal{L}_n^{\mathrm{FR}}\big(S_n, \Sigma_n\big)\big] - \mathbb{E}\big[\mathcal{L}_n^{\mathrm{FR}}\big(S_n^*, \Sigma_n\big)\big]} \times 100\% \ . \tag{B.1}$$

B.1 Concentration Ratio

First we revisit the results of Section 6.5, where the concentration ratio varies while the other simulation parameters remain fixed as per the baseline scenario. The equivalent to

Figure 6 in terms of the Frobenius loss is Figure 10 below.

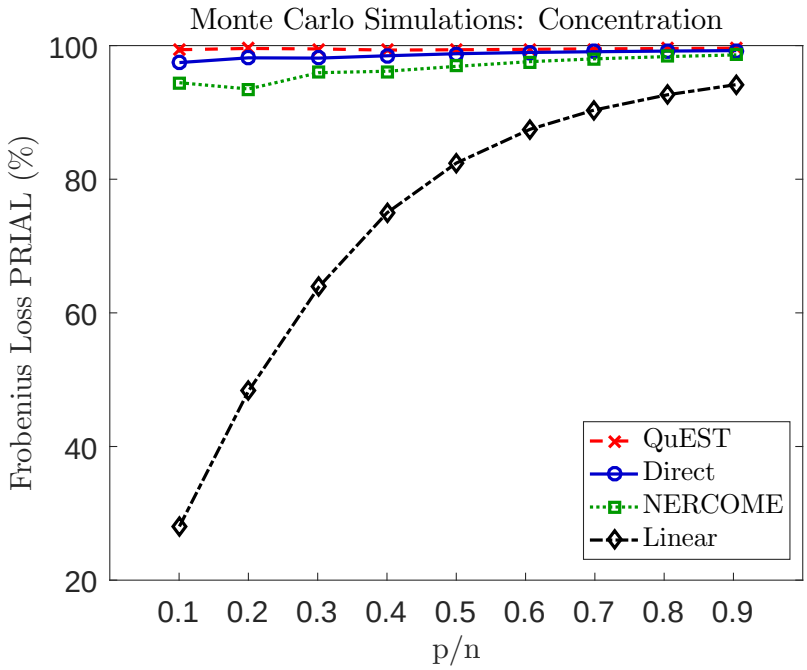

Figure 10: Evolution of the Frobenius PRIAL of various estimators as a function of the ratio of matrix dimension to sample size.

Linear shrinkage is now comfortably above the sample covariance matrix. This confirms that any underperformance observed in Section 6.5 is solely attributable to the choice of a loss function that is the 'wrong' one for linear shrinkage.

B.2 Condition Number

Second we revisit the results of Section 6.6, where the condition number varies from $\theta = 3$ to $\theta = 30$ while the other simulation parameters remain fixed as per the baseline scenario. The equivalent to Figure 7 in terms of the Frobenius loss is Figure 11 below.

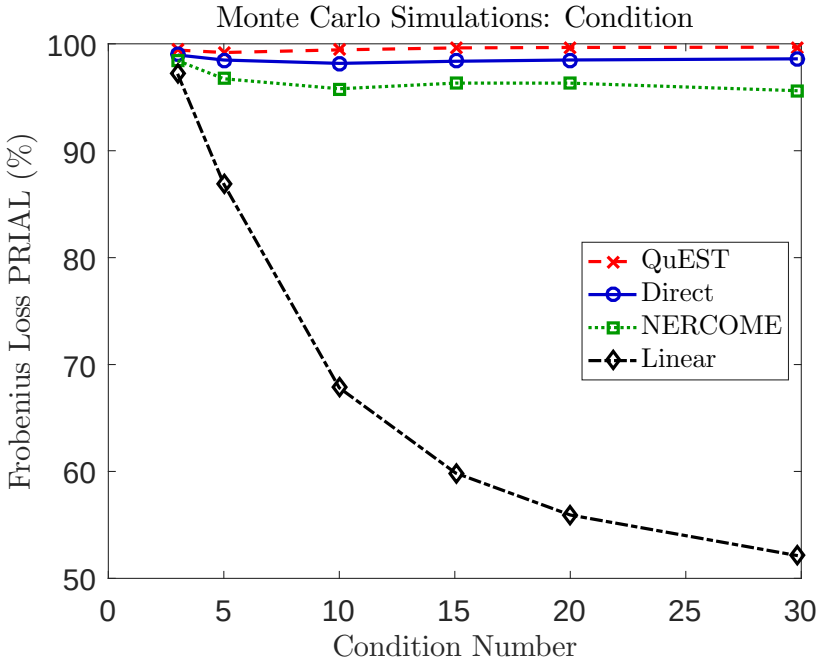

Figure 11: Evolution of the Frobenius PRIAL of various estimators as a function of the condition number of the population covariance matrix.

Linear shrinkage is also comfortably above the sample covariance matrix. Any underperformance observed in Section 6.5 is solely attributable to the loss function.

B.3 Fixed-Dimension Asymptotics

Third and last, we revisit the results of Section 6.9, where the sample size goes from $n = 250$ to $n = 20,000$ while all the other simulation parameters remain fixed as per the baseline scenario. The equivalent to Figure 8 in terms of the Frobenius loss is Figure 12 below.

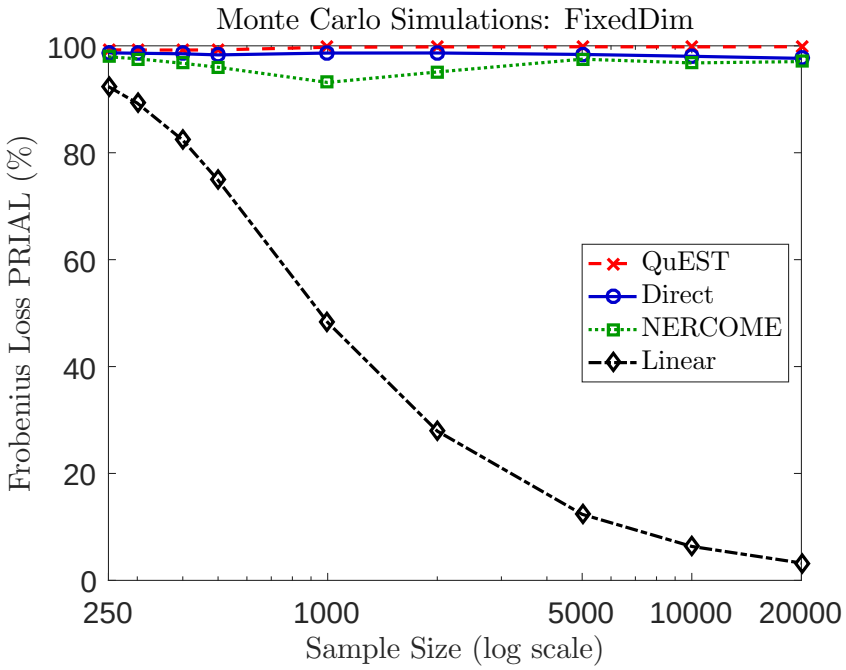

Figure 12: Evolution of the Frobenius PRIAL of various estimators as the sample size grows towards infinity, while the matrix dimension remains fixed.

Linear shrinkage again improves over the sample covariance matrix.

B.4 Overall Assessment

The ultimate conclusion of this investigation is that any weakness of linear shrinkage relative to the sample covariance matrix in terms of the minimum variance loss is solely due to the fact that the linear shrinkage estimator of Ledoit and Wolf (2004) is based on the Frobenius loss instead.

C Singular Case

When the matrix dimension p exceeds the sample size n, the $p-n$ smallest eigenvalues $(\lambda_1, \ldots, \lambda_{p-n})$ are all equal to zero. Thus, the attention shifts away from the sample spectral e.d.f. F_n towards the e.d.f. of the n *nonzero* sample eigenvalues, which is defined as

$$\forall x \in \mathbb{R} \qquad \underline{F}_n(x) := \frac{1}{n} \sum_{i=p-n+1}^{p} \mathbb{1}_{\{x \geq \lambda_{n,i}\}} \ .$$

The function $\underline{F}_n$ is the spectral e.d.f. of the matrix $Y_n Y_n'/n = X_n \Sigma_n X_n'/n$. The relationship between the two e.d.f.'s is

$$\forall x \in \mathbb{R} \qquad F_n(x) = \frac{p-n}{p} \mathbb{1}_{\{x \geq 0\}} + \frac{n}{p} \underline{F}_n(x) \ .$$

When $c \in (1, +\infty)$, there exists a limiting c.d.f. $\underline{F}$ such that $\forall x \in \mathbb{R}$ $\underline{F}_n(x) \xrightarrow{\text{a.s.}} \underline{F}(x)$. The limiting c.d.f. $\underline{F}$ admits a continuous derivative $\underline{f}$ on $\mathbb{R}$. Its Hilbert transform $\mathcal{H}_{\underline{f}}$ also exists and is continuous. The following relationships hold:

$$\forall x \in (0, +\infty) \qquad \underline{f}(x) = cf(x) \tag{C.1}$$

$$\mathcal{H}_{\underline{f}}(x) = \frac{c-1}{\pi x} + c\mathcal{H}_f(x) \ . \tag{C.2}$$

All of this follows directly from Silverstein (1995) and Silverstein and Choi (1995), if we replace $c \in (0, 1)$ with $c \in (1, +\infty)$ in Assumption 1. The oracle nonlinear shrinkage function defined in Equation (3.2) can be rewritten in terms of these new objects as

$$d^{\text{o}}(x) = \frac{x}{\pi^2 x^2 \left[\underline{f}(x)^2 + \mathcal{H}_{\underline{f}}(x)^2\right]} \ . \tag{C.3}$$

A similar formulation is attained in Equation (8) of Ledoit and Wolf (2017a).

We adapt the kernel method developed in Section 4 to estimate the limiting density $\underline{f}$ with

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

and its Hilbert transform $\mathcal{H}_{\underline{f}}$ with

$$\forall x \in \mathbb{R} \qquad \mathcal{H}_{\widetilde{\underline{f}}_n}(x) := \frac{1}{n} \sum_{i=p-n+1}^{p} \frac{1}{h_{n,i}} \mathcal{H}_k\left(\frac{x - \lambda_{n,i}}{h_{n,i}}\right) = PV \int \frac{\widetilde{\underline{f}}_n(t)}{x - t} dt \ .$$

From these two estimators we deduce the shrunk eigenvalues in a manner analogous to Equation (4.3):

$$\forall i = p-n+1, \ldots, n \qquad \widetilde{d}_{n,i} := \frac{\lambda_{n,i}}{\pi^2 \lambda_{n,i}^2 \left[\widetilde{\underline{f}}_n(x)^2 + \mathcal{H}_{\widetilde{\underline{f}}_n(x)^2}\right]} \ . \tag{C.4}$$

The only question remaining is how to handle the null eigenvalues $(\lambda_1, \ldots, \lambda_{p-n})$. Theorem 2 of Ledoit and Wolf (2017a), building on Equation (13) of Ledoit and Péché (2011), shows that the oracle shrinkage formula is a different one, namely,

$$d^{\text{o}}(0) := \frac{1}{\pi(c-1)\mathcal{H}_{\underline{f}}(0)} \ .$$

In keeping with the procedure adopted so far, we estimate it with

$$\forall i = 1, \ldots, p-n \qquad \widetilde{d}_{n,i} := \frac{1}{\pi \dfrac{p-n}{n} \mathcal{H}_{\widetilde{\underline{f}}_n}(0)} \ . \tag{C.5}$$

As before, we operationalize these formulas with the semicircular kernel $\kappa(x)$ and the proportional bandwidth $h_{n,i} := \lambda_{n,i} h_n$:

$$\widetilde{\underline{f}}_n(\lambda_{n,i}) = \frac{1}{n} \sum_{j=p-n+1}^{p} \frac{\sqrt{\left[4\lambda_{n,j}^2 h_n^2 - (\lambda_{n,i} - \lambda_{n,j})^2\right]^+}}{2\pi\lambda_{n,j}^2 h_n^2} \qquad \forall i = p-n+1, \ldots, p \tag{C.6}$$

$$\mathcal{H}_{\widetilde{\underline{f}}_n}(\lambda_{n,i}) = \frac{1}{n} \sum_{j=p-n+1}^{p} \frac{\operatorname{sgn}(\lambda_{n,i} - \lambda_{n,j})\sqrt{\left[(\lambda_{n,i} - \lambda_{n,j})^2 - 4\lambda_{n,j}^2 h_n^2\right]^+} - \lambda_{n,i} + \lambda_{n,j}}{2\pi\lambda_{n,j}^2 h_n^2} \,. \tag{C.7}$$

$$\mathcal{H}_{\widetilde{\underline{f}}_n}(0) = \frac{1 - \sqrt{1 - 4h_n^2}}{2\pi n h_n^2} \sum_{j=p-n+1}^{p} \frac{1}{\lambda_{n,j}} = \frac{2}{1 + \sqrt{1 + 4h_n^2}} \mathcal{H}_{\underline{F}'_n}(0) \,. \tag{C.8}$$

Note that, as $n \to \infty$, the multiplier in front of $\mathcal{H}_{\underline{F}'_n}(0)$ in (C.8) goes to one, so $\mathcal{H}_{\widetilde{\underline{f}}_n}(0)$ and $\mathcal{H}_{\underline{F}'_n}(0)$ are asymptotically equivalent.

Finally, we regroup the nonlinearly shrunk eigenvalues from (C.4) and (C.5) into the vector $\widetilde{\boldsymbol{d}}_n$, apply the isotonization algorithm $\widehat{\boldsymbol{d}}_n = \mathrm{PAV}\big(\widetilde{\boldsymbol{d}}_n\big)$, and recompose with sample eigenvectors to compute the covariance matrix estimator $\widehat{S}_n = \sum_{i=1}^{p} \widehat{d}_{n,i} \cdot u_{n,i} u_{n,i}'$.

This enables us to run a counterpart of the Monte Carlo simulations in Section 6.5 for the case $c > 1$. We vary the concentration ratio p/n from 1.1 to 10 while holding the product $p \times n$ constant at the level it had under the baseline scenario, namely, $p \times n = 120,000$. The PRIALs are displayed in Figure 13. Given that the minimum variance loss $\mathcal{L}_n^{\mathrm{MV}}$ of the sample covariance matrix is undefined, due to S_n being singular in this case, we report the PRIAL with respect to the Frobenius loss $\mathcal{L}_n^{\mathrm{FR}}$ instead, as in Appendix B. Qualitatively, there is no difference across the two loss functions in terms of rankings between estimators and proximity to the ideal FSOPT benchmark.

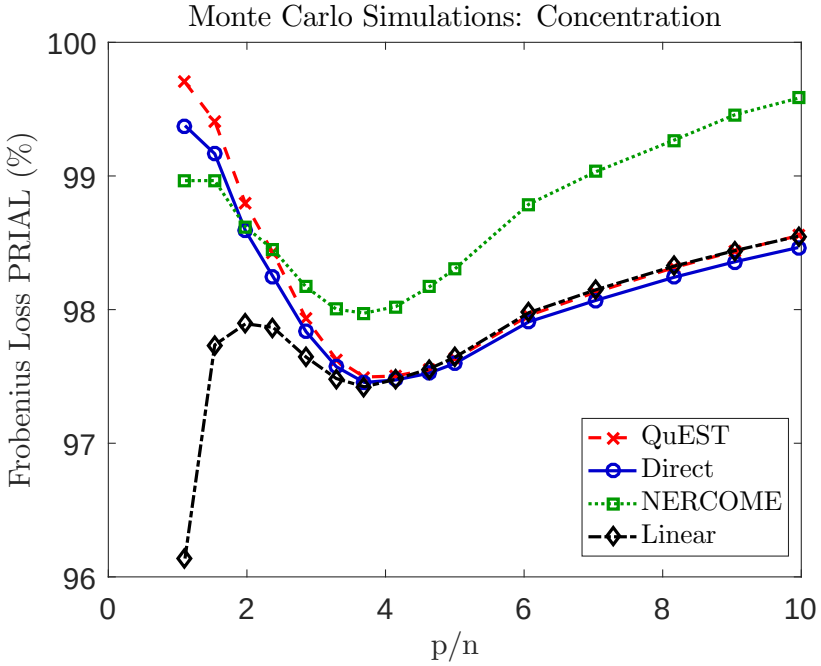

Figure 13: Evolution of the Frobenius PRIAL of various estimators when the matrix dimension exceeds the sample size.

We draw the attention of the reader to the vertical scale of the figure: It starts at 96%. This confirms the trend that could be inferred from Figure 6: Higher concentration ratios make all shrinkage estimators look good. At this level of performance, the exact ordering becomes relatively less important, but NERCOME seems to pull ahead of the pack for $p/n > 2$, perhaps due to the fact that the sample size n is not so big anymore.

D Programming Code

The Matlab function that computes the nonlinear shrinkage estimator of the covariance matrix based on our new methodology has only 20 lines of actual Matlab code, which makes for easy debugging and customization.

```
function sigmahat=direct_kernel(X)
% extract sample eigenvalues sorted in ascending order and eigenvectors
[n,p]=size(X);
sample=(X'*X)./n;
[u,lambda]=eig(sample,'vector');
[lambda,isort]=sort(lambda);
u=u(:,isort);
% compute direct kernel estimator
lambda=lambda(max(1,p-n+1):p);
L=repmat(lambda,[1 min(p,n)]);
h=n^(-0.35);                                              % Equation (5.4)
ftilde=mean(sqrt(max(0,4*L'.^2*h^2-(L-L').^2))./(2*pi*L'.^2*h^2),2);%(5.2)
Hftilde=mean((sign(L-L').*sqrt(max(0,(L-L').^2-4*L'.^2*h^2))-L+L') ...
   ./(2*pi*L'.^2*h^2),2);                                 % Equation (5.3)
if p<=n
   dtilde=lambda./((pi*(p/n)*lambda.*ftilde).^2 ...
      +(1-(p/n)-pi*(p/n)*lambda.*Hftilde).^2);            % Equation (4.3)
else
   Hftilde0=(1-sqrt(1-4*h^2))/(2*pi*h^2)*mean(1./lambda); % Equation (C.8)
   dtilde0=1/(pi*(p-n)/n*Hftilde0);                       % Equation (C.5)
   dtilde1=lambda./(pi^2*lambda.^2.*(ftilde.^2+Hftilde.^2));   % Eq. (C.4)
   dtilde=[dtilde0*ones(p-n,1);dtilde1];
end
dhat=pav(dtilde);                                         % Equation (4.5)
sigmahat=u*(repmat(dhat,[1 p]).*u');                      % Equation (4.6)
```

The direct kernel function transforms an $n \times p$ matrix `X` containing n iid samples of p variables into the $p \times p$ nonlinear shrinkage covariance matrix estimator `sigmahat`. It invokes the `pav` function from the `math` subfolder of the freely available E-MAP toolbox version 2.0 at http://sourceforge.net/projects/emap-toolbox/, which was originally developed by Scott R. Collins for large-scale genomics; see Collins et al. (2006).

References

Abadir, K., Distaso, W., and Žikešš, F. (2014). Design-free estimation of variance matrices. *Journal of Econometrics*, 181:165–180.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., Silverman, E., et al. (1955). An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics*, 26(4):641–647.

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

Capon, J. (1969). High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, 57(8):1408–1418.

Collins, S. R., Schuldiner, M., Krogan, N. J., and Weissman, J. S. (2006). A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. *Genome Biology*, 7(7):R63.

Engle, R. F. and Colacito, R. (2006). Testing and valuing dynamic correlations for asset allocation. *Journal of Business and Economic Statistics*, 24(2):238–253.

Engle, R. F., Ledoit, O., and Wolf, M. (2017). Large dynamic covariance matrices. *Journal of Business & Economic Statistics*. doi: 0.1080/07350015.2017.1345683.

Erdélyi, A., editor (1954). *Tables of Integral Transforms*, volume II of *California Institute of Technology, Bateman Manuscript Project*. McGraw-Hill, New York. Based, in part, on notes left by Harry Bateman.

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. *Annals of Statistics*, 38(6):3724–3750.

Krantz, S. G. (2009). *Explorations in Harmonic Analysis*. Birkhäuser, Boston.

Lam, C. (2016). Nonparametric eigenvalue-regularized precision or covariance matrix estimator. *The Annals of Statistics*, 44(3):928–953.

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. (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. (2017a). Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets Goldilocks. *Review of Financial Studies*. doi: 10.1093/rfs/hhx052.

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

Ledoit, O. and Wolf, M. (2017c). Optimal estimation of a large-dimensional covariance matrix under Stein's loss. *Bernoulli*. Forthcoming.

Marčenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Sbornik: Mathematics*, 1(4):457–483.

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

Rajaratnam, B. and Vincenzi, D. (2016). A theoretical study of stein's covariance estimator. *Biometrika*, 103(3):653–666.

Ribes, A., Azas, J.-M., and Planton, S. (2009). Adaptation of the optimal fingerprint method for climate change detection using a well-conditioned covariance matrix estimate. *Climate Dynamics*, 33(5):707–722.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, Boca Raton.

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 Bai, Z. D. (1995). On the empirical distribution of eigenvalues of a class of large-dimensional random matrices. *Journal of Multivariate Analysis*, 54:175–192.

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

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

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