University of Zurich

Department of Economics

Working Paper Series

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

Working Paper No. 264

Analytical Nonlinear Shrinkage of Large-Dimensional Covariance Matrices

Olivier Ledoit and Michael Wolf

First version: September 2017
This version: November 2018

Analytical Nonlinear Shrinkage 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

First version: September 2017
This version: November 2018

Abstract

This paper establishes the first analytical formula for optimal nonlinear shrinkage of large-dimensional covariance matrices. We achieve this by identifying and mathematically exploiting a deep connection between nonlinear shrinkage and nonparametric estimation of the Hilbert transform of the sample spectral density. Previous nonlinear shrinkage methods were numerical: QuEST requires numerical inversion of a complex equation from random matrix theory whereas NERCOME is based on a sample-splitting scheme. The new analytical approach is more elegant and also has more potential to accommodate future variations or extensions. Immediate benefits are that it is typically 1,000 times faster with the same accuracy, and accommodates covariance matrices of dimension up to $10,000$. The difficult case where the matrix dimension exceeds the sample size is also covered.

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, p, are commensurate with the sample size, n, attention is naturally turning to the more difficult — but potentially more rewarding — method of *nonlinear* shrinkage estimation, where the transformation applied to the sample eigenvalues must be optimal not in a space of dimension two (intercept and slope) but in a much larger space of dimension p (that is, unconstrained transformation).

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

The second method, going back to Abadir 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 be more accurate or faster, but in the fact that it is decisively simpler and more transparent, thus providing an independent and easily verifiable confirmation for the mathematically delicate *indirect* method of QuEST.

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

We contribute to the existing literature on three levels. At the conceptual level, we show how the presence of the Hilbert transform in the shrinkage formula is the ingredient that induces "shrinkage" by attracting nearby eigenvalues 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; indeed, from the point of view of optimal covariance matrix estimation, the Hilbert transform is equally as important as the density itself. Krantz (2009, p. 17) alludes to this importance being commonplace in mathematics: "The Hilbert transform is, without question, the most important operator in analysis. It arises in many different contexts, and all these contexts are intertwined in profound and influential ways." Our second extension of the kernel estimator is that, instead of keeping the bandwidth constant (or uniform) for a given sample size, we let it vary in proportion to the location of a given sample eigenvalue. This improvement confines the support of the spectral density estimator to the positive half of the real line, as befits positive-definite matrices.; it also reflects the scale-invariance of the problem. Finally, at the operational level, we make the computer code two orders of magnitude simpler and faster than the 'indirect' route of numerically inverting the QuEST function. As a result, we can estimate covariance matrices of dimension $10,000$ and beyond, whereas the largest magnitude that could be handled by nonlinear shrinkage before was $1,000$.

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 runs an extensive set of Monte Carlo simulations. Section 6 studies the robustness of the performance of the analytical nonlinear shrinkage estimator against alternative choices of kernel and bandwidth. Section 7 concludes. An appendix contains all mathematical proofs, further simulation results, the extension to the singular case, and our code.

2 Finite Samples

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

2.1 Rotation Equivariance

Let Σ_n denote a p-dimensional population covariance matrix. A mean-zero independent and identically distributed (i.i.d.) i.i.d. sample of n observations Y_n generates the sample

covariance matrix $S_n \coloneqq Y_n'Y_n/n$. 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 $\widehat{\Sigma}_n \coloneqq U_n\widehat{\Delta}_n U_n'$, where $\widehat{\Delta}_n$ is a diagonal matrix whose elements $\widehat{\boldsymbol{\delta}}_n = \big(\widehat{\delta}_{n,1}, \ldots, \widehat{\delta}_{n,p}\big) \in (0, +\infty)^p$ are a function of $\boldsymbol{\lambda}_n$. Thus, $\widehat{\Sigma}_n = \sum_{i=1}^p \widehat{\delta}_{n,i} \cdot u_{n,i}u_{n,i}'$.

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

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 1). It can be expressed in our notation as

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

$\mathcal{L}_n^{\text{MV}}$ 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 $\widehat{\Sigma}_n \coloneqq \sum_{i=1}^p \widehat{\delta}_{n,i} \cdot u_{n,i}u_{n,i}'$ 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, finite-sample optimal nonlinear shrinkage replaces the sample eigenvalues $\boldsymbol{\lambda}_n$ with the unobservable quantity

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

prior to recombining it with the sample eigenvectors to form the (non-feasible) covariance matrix estimator

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

Remark 2.1. Section 3.1 of Ledoit and Wolf (2012) shows that the same estimator S_n^* is also finite-sample optimal with respect to the (squared) Frobenius loss function, which is defined for generic estimator $\widehat{\Sigma}_n$ as

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

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

3 Large-Dimensional Asymptotics

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

3.1 Assumptions

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

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

The case $c > 1$, where the sample covariance matrix is singular, is covered in Appendix C.

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 $x \longmapsto \sum_{i=1}^{p} \mathbb{1}_{\{\alpha_i \leq x\}}/p$, where $\mathbb{1}$ denotes the indicator function.*

Assumption 3.2 (Population Covariance Matrix)**.**

a. *The population covariance matrix Σ_n is a nonrandom symmetric positive-definite matrix of dimension $p \times p$.*
b. *Let $\boldsymbol{\tau}_n := (\tau_{n,1}, \ldots, \tau_{n,p})'$ 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 3.3 (Data Generating Process)**.** *X_n is an $n \times p$ matrix of i.i.d. random variables with mean zero, variance one, and finite 16th moment. The matrix of observations is $Y_n := X_n \times \sqrt{\Sigma_n}$. Neither $\sqrt{\Sigma_n}$ nor X_n are observed on their own: only Y_n is observed.*

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

The sample covariance matrix S_n, its eigenvalues $\boldsymbol{\lambda}_n := (\lambda_{n,1}, \ldots, \lambda_{n,p})$ and eigenvectors $U_n := [u_{n,1} \ldots u_{n,p}]$ have already been defined in Section 2.1. The e.d.f. of sample eigenvalues is the function $F_n(x) := \sum_{i=1}^{p} \mathbb{1}_{\{x \geq \lambda_{n,i}\}}/p$ 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 3.1–3.3, there exists a limiting sample spectral distribution F such that $\forall x \in \mathbb{R}\ F_n(x) \xrightarrow{\text{a.s.}} F(x)$. The limiting sample spectral c.d.f. F is uniquely determined by c and H; therefore, we will refer to it as $F_{c,H} := F$ whenever clarification is needed.

Assumptions 3.1–3.3 together with Theorem 1.1. of Bai and Silverstein (1998) imply that the support of F, denoted by $\mathsf{Supp}(F)$, is the union of a finite number $\nu \geq 1$ 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* $\frac{dt}{\pi t}$.

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

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

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

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

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

The intuition behind the Hilbert transform is that it operates like a local attraction force. It is very positive if there are heavy mass points slightly larger than you, so it pushes you up (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 visual inspection of the Cauchy kernel. Figure 1 confirms them visually by plotting the Hilbert transform of four well-known densities.

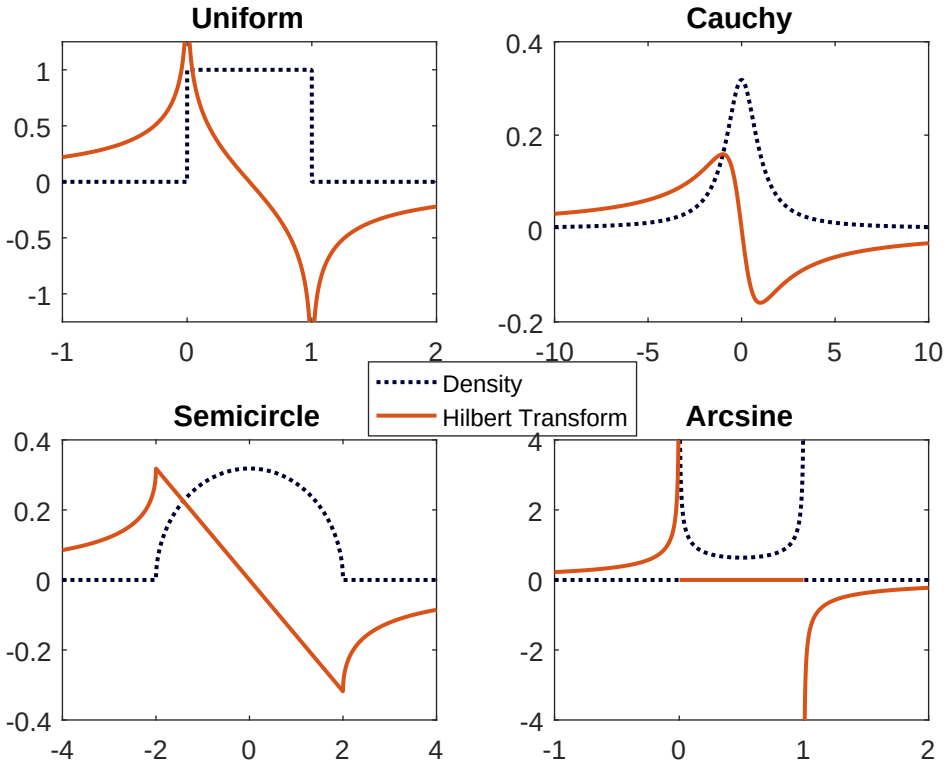

Figure 1: Hilbert transform of four densities. The transform is strongly positive to the left of the center of mass, strongly negative to the right, and vanishes away from the center of mass.

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

	Density	Hilbert Transform
Uniform	$f(x) = \mathbb{1}_{\{0 \leq x < 1\}}$	$\mathcal{H}_f(x) = \frac{1}{\pi} \log \left\lvert \frac{1-x}{x} \right\rvert$
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 two 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 3.4 (Limiting Shrinkage Function)**.** *There exists a nonrandom real univariate function* $\widehat{\delta}$ *defined on* $\mathsf{Supp}(F)$ *and continuously differentiable such that* $\widehat{\delta}_n(x) \xrightarrow{\text{a.s.}} \widehat{\delta}(x)$*, for all* $x \in \mathsf{Supp}(F)$*. Furthermore, this convergence is uniform over* $x \in \bigcup_{k=1}^{\nu} [a_k + \eta, b_k - \eta]$*, for any small* $\eta > 0$*. Finally, for any small* $\eta > 0$*, there exists a*

finite nonrandom constant $\widehat{K}$ *such that almost surely, over the set* $x \in \bigcup_{k=1}^{\nu}[a_k - \eta, b_k + \eta]$, $\widehat{\delta}_n(x)$ *is uniformly bounded by* $\widehat{K}$ *from above and by* $1/\widehat{K}$ *from below, for all* n *large enough.*

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

Theorem 3.1. *Define the oracle nonlinear shrinkage function*

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

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

(a) The oracle estimator of the covariance matrix

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

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

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

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

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

Remark 3.2. The quantities $\big(d^{\mathrm{o}}(\lambda_{n,1}), \ldots, d^{\mathrm{o}}(\lambda_{n,p})\big)$ represent large-dimensional asymptotic counterparts to the finite-sample optimal quantities $\big(u_{n,1}' \Sigma_n u_{n,1}, \ldots, u_{n,p}' \Sigma_n u_{n,p}\big)$ of Equation (2.2). Equation (3.3) was first discovered by Ledoit and Péché (2011, Theorem 3), based on a generalization of the Fundamental Equation of Random Matrix Theory originally due to Marčenko and Pastur (1967). The formula here is the first one expressed without any reference to complex numbers; previous (mathematically equivalent) versions used the complex-valued Stieltjes transform instead of the Hilbert transform. ■

3.5 Shrinkage as Local Attraction via the Hilbert Transform

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

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

which is easier to interpret. If the limiting concentration ratio c is negligible, then the denominator is close to one, which would mean no shrinkage: This is why the sample covariance matrix works well under traditional (fixed-dimensional) asymptotics. As c increases, however, (noticeable) shrinkage must occur. Let us set aside the term $c^2\left[\varphi(x)^2 + H_\varphi(x)^2\right]$ 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, which cannot happen with linear shrinkage instead: Any eigenvalue below average will be brought up necessarily. Figure 2 provides a graphical illustration for these contrasting behaviors.

In this example, the average eigenvalue is equal to 1.25. Sample eigenvalues below the average but above 1 need to be "shrunk" 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

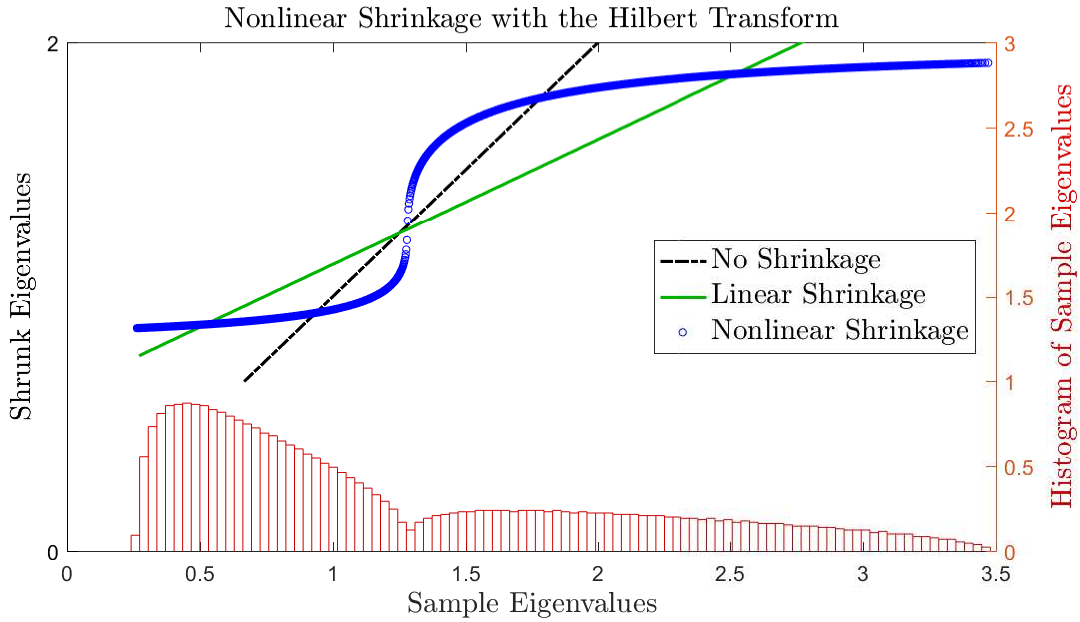

Figure 2: Local Attraction Effect. $2,500$ population eigenvalues are equal to 0.8, and $1,500$ are equal to 2. The sample size is $n = 18,000$. At the bottom of the figure is a histogram displaying the location of the sample eigenvalues.

their immediate right. Linear shrinkage, being a global operator, is not equipped to sense a disturbance in the force: It applies the same shrinkage intensity across the board and shrinks all eigenvalues towards the average of 1.25.

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 is feasible in practice) with the same asymptotic properties.

Remark 3.3. Ledoit and Wolf (2018, Section 4.2) prove that the estimator S_n^{o} is also optimal within the class of rotation-equivariant estimators of Assumption 3.4 with respect to the Frobenius norm. Intuitively, this is because the two corresponding finite-sample optimal estimators are identical, as pointed out in Remark 2.1. ■

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 formula (3.3). Things are more complicated,

however, for the limiting sample spectral density f and its Hilbert transform $\mathcal{H}_f$. Given that the sample spectral e.d.f F_n converges to F almost surely, the obvious idea would have been to plug its derivative F_n' in place of f:

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

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

4 Asymptotic Theory

4.1 Kernel Requirements

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

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

4.2 Proportional Bandwidth

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

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

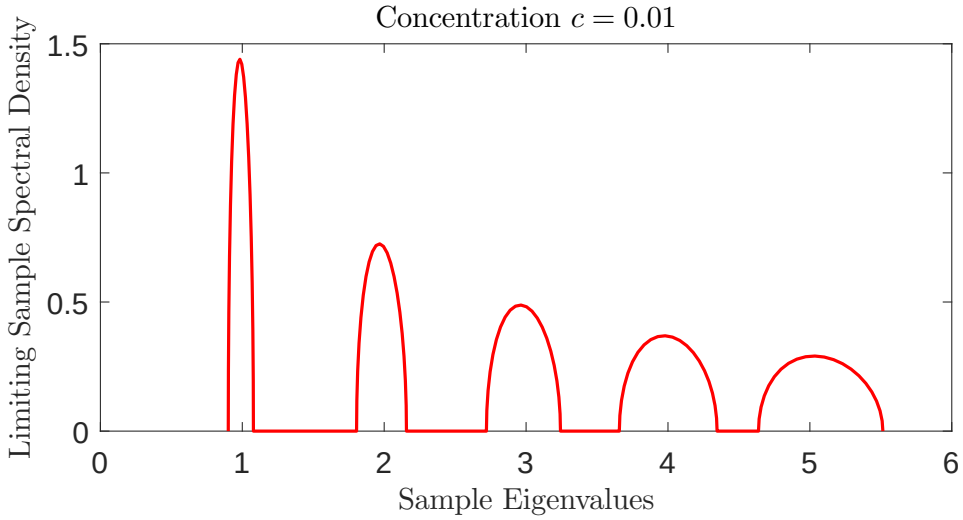

Figure 3: Limiting sample spectral density f when the population eigenvalues are $\{1, 2, \ldots, 5\}$, each with weight $1/5$.

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

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

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

4.3 Kernel Estimators

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

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

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

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

4.4 Uniform Consistency

Our main results are as follows. All proofs are in Appendix A.

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

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

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

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

in probability.

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

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

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

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

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

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

Remark 4.1. The above statement remains true if we replace the minimum variance loss function $\mathcal{L}_n^{\text{MV}}$ with the Frobenius loss function $\mathcal{L}_n^{\text{FR}}$. Indeed, Section 4.2 of Ledoit and Wolf (2018) proves that the same estimator S_n^{o} is also optimal within the class of rotation-equivariant estimators of Assumption 3.4 with respect to the Frobenius norm. Intuitively, this is because the finite-sample optimal estimator of Equation (2.2) is the same as the finite-sample optimal estimator with respect to the Frobenius norm, given in Ledoit and Wolf (2012, Equation (3.2)). ■

4.5 Epanechnikov Kernel

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

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

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

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

Computation The Hilbert transform of the Gaussian kernel is the Dawson (1897) integral, which is a higher-transcendental function extremely slow to compute.

The kernel originally introduced by Epanechnikov (1969) as per his Equation (13) has unit variance, support $\left[-\sqrt{5}, \sqrt{5}\right]$, and density

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

This is the default kernel used for univariate density estimation by the popular software STATA, among others. Its Hilbert transform does not appear to have been computed in the literature. We derive it in the following proposition.

Proposition 4.1.

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

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

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

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

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

Throughout, the last term in (4.8) is understood to be a zero in the unlikely event that $(\lambda_{n,i} - \lambda_{n,j})^2$ happens to be exactly equal to $5\lambda_{n,j}^2h_n^2$. Alternative kernels are explored as robustness checks through Monte Carlo simulations in Section 6.

4.6 Choice of Bandwidth

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

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

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

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

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

has support is $\left[\lambda_{n,i} \left(1 - \sqrt{5} n^{-1/3} \right) , \lambda_{n,i} \left(1 + \sqrt{5} n^{-1/3} \right) \right]$. The lower boundary is in the positive half-line if and only if $n > 5\sqrt{5} \approx 11.2$, therefore it is unadvisable to use this large-dimensional asymptotic procedure when $p < 12$. Alternative choices of the bandwidth are explored as robustness checks through Monte Carlo simulations in Section 6.

4.7 Summary: The Analytical Formula

Compute the spectral decomposition of the sample covariance matrix as per Section 2.1:

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

Choose the global bandwidth as per Section 4.6:

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

Specify the locally adaptive bandwidth as per Section 4.2:

$$\forall j = 1, \ldots, n \qquad h_{n,j} := \lambda_{n,j} h_n \ .$$

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

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

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

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

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

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

Recompose the covariance matrix estimator as per Section 4.4:

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

Computer code for this analytical formula is in Appendix D.

5 Monte Carlo Simulations

5.1 Competitors

We compare the performance of the following six covariance matrix estimators:

Sample The sample covariance matrix S_n.

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

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

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

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

FSOPT The finite-sample optimal estimator S_n^* defined in Equation (2.3), which would require knowledge of the unobservable population covariance matrix Σ_n, and thus is not applicable in the real world but serves as 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. The code for the QuEST package comes from the numerical implementation detailed in Ledoit and Wolf (2017b) and is freely downloadable from the academic website of the second author.

5.2 Percentage Relative Improvement in Average Loss

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

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

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

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

Also by construction $\text{PRIAL}_n^{\text{MV}}(S_n^*) = 100\%$ because this is the maximum amount of loss reduction that can be attained by nonlinear shrinkage within the rotation-equivariant framework of Section 2.1, as shown in Proposition 2.1. Given that the construction of S_n^* requires knowledge of the 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.

5.3 Baseline Scenario

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

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

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

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

Estimator	Sample	Linear	Analytical	QuEST	NERCOME	FSOPT
Average Loss	2.71	2.10	1.52	1.50	1.58	1.48
PRIAL	0%	50%	97%	98%	92%	100%
Time (ms)	1	2	3	2,346	3,071	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/Analytical/QuEST), which is a very satisfactory number.

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

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

Table 2 demonstrates that applied researchers who are already comfortable with linear shrinkage and would like to upgrade to nonlinear shrinkage for performance enhancement, but have been concerned by the numerical complexity of the earlier techniques, can now safely upgrade to the analytical formula.

5.4 Convergence

5.4.1 Large-Dimensional Asymptotic Performance

Under large-dimensional asymptotics, the matrix dimension p and the sample size n go to infinity together, while their ratio p/n converges to some limit c. In the first experiment,

we let p and n vary together, with their ratio fixed at the baseline value of 1/3. The results are displayed in Figure 4.

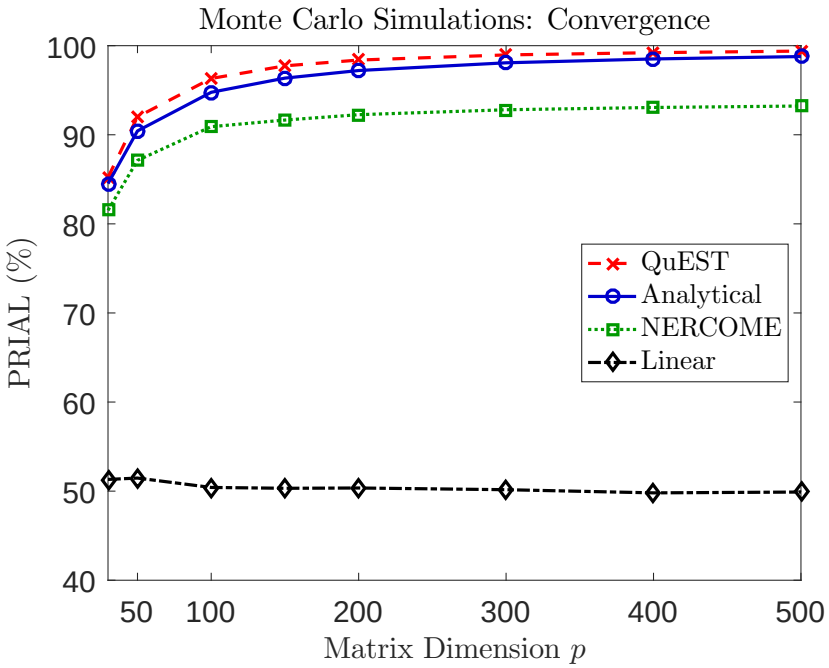

Figure 4: Evolution of the PRIAL of various estimators as the matrix dimension and the sample size go to infinity together.

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 Analytical is never more than 2%, which is very small.

5.4.2 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.

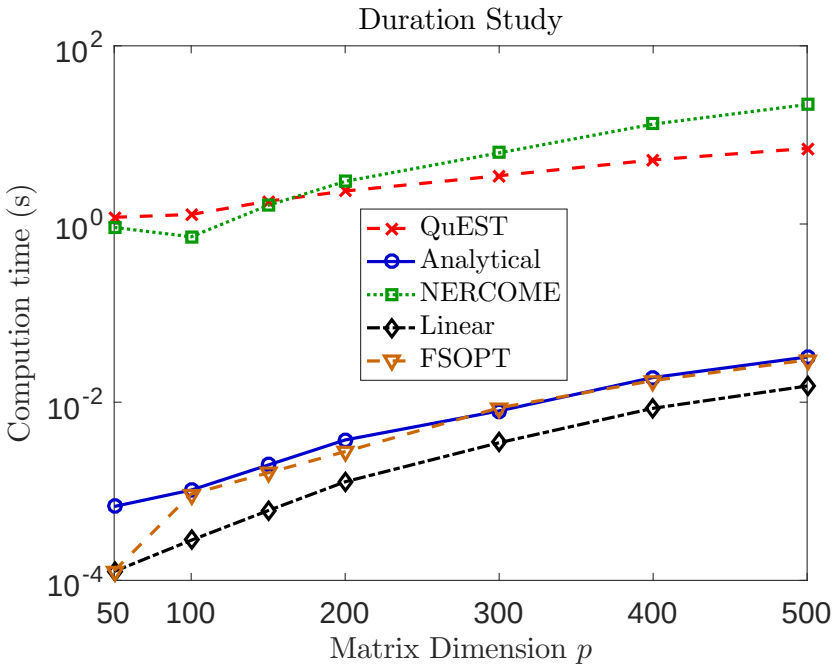

Figure 5: Computational speed of various shrinkage estimators as the matrix dimension and the sample size go to infinity together, measured in seconds, with log-scale on the vertical axis.

There is a clear gap between, on the one hand, QuEST and NERCOME and, on the other hand, Analytical, Linear and FSOPT. Analytical nonlinear shrinkage is typically 1,000 times faster than its numerical counterparts.

5.4.3 Ultra-Large 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	Analytical	FSOPT
Average Loss	2.679	2.086	1.488	1.487
PRIAL	0%	49.74%	99.90%	100%
Time (s)	21	43	113	108

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 the analytical nonlinear shrinkage estimator gets ever closer to 100%, as expected from theory.

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

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

5.5 Concentration Ratio

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

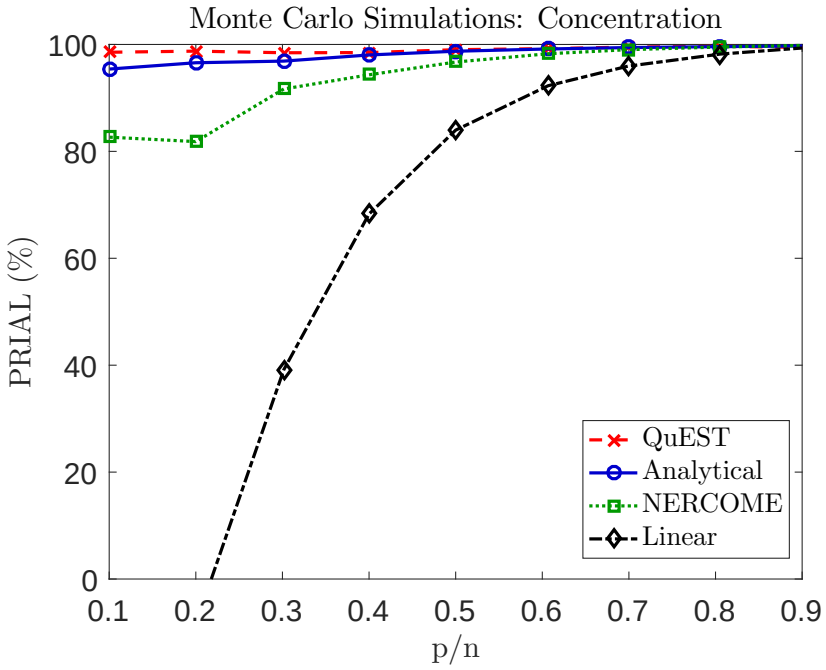

Figure 6: Evolution of the PRIAL of various estimators as a function of the ratio of the matrix dimension to the sample size.

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 finding is due to the fact that linear shrinkage is optimized for a different loss function than the minimum variance loss, namely, the Frobenius loss. Under Frobenius loss, linear shrinkage always beats the sample covariance matrix in the same simulation experiment.

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

5.6 Condition Number

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

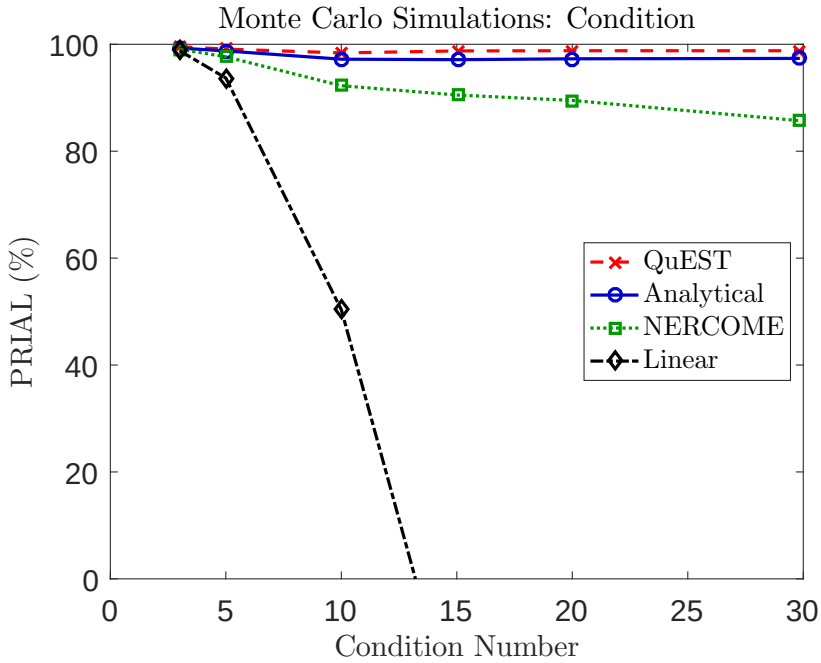

Figure 7: Evolution of the PRIAL of various estimators as a function of the condition number of the population covariance matrix.

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 Analytical in particular being very close to QuEST and above the 97% mark across the board.

5.7 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	Analytical	QuEST	NERCOME
Bernoulli	51%	97%	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.

5.8 Shape of the Distribution of Population Eigenvalues

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

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

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

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

5.9 Fixed-Dimensional Asymptotics

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

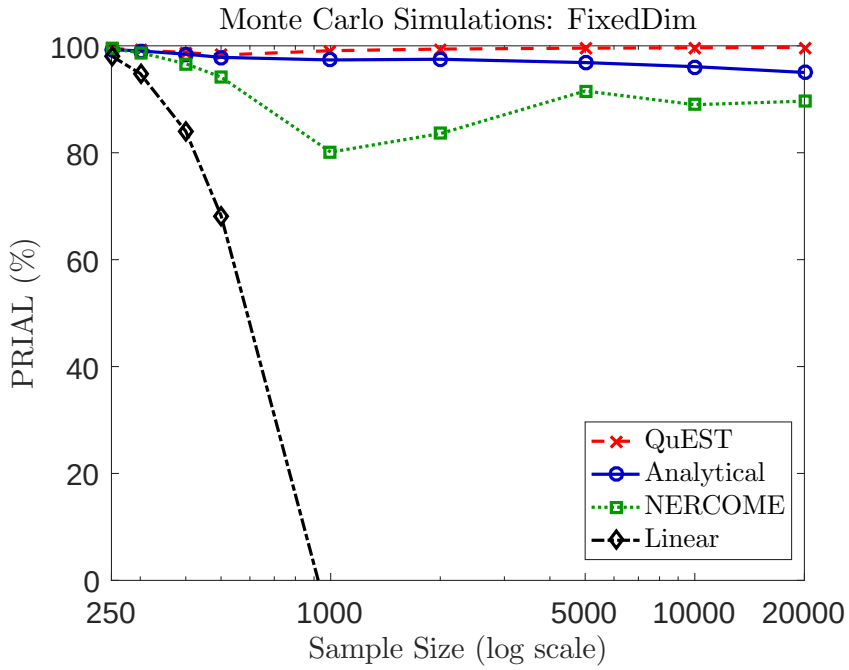

Figure 8: Evolution of the PRIAL as the sample size grows towards infinity, while the matrix dimension remains fixed.

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 Analytical in particular being very close to QuEST and above the 96% mark across the board.

5.10 Arrow Model

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

Inspired by a factor model where all pairs of variables have 50% correlation and all variables have unit standard deviation, and by the 'arrow model' introduced by Ledoit and Wolf (2018, Section 7), we set the largest eigenvalue (the 'arrow' 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 $p = 500$.

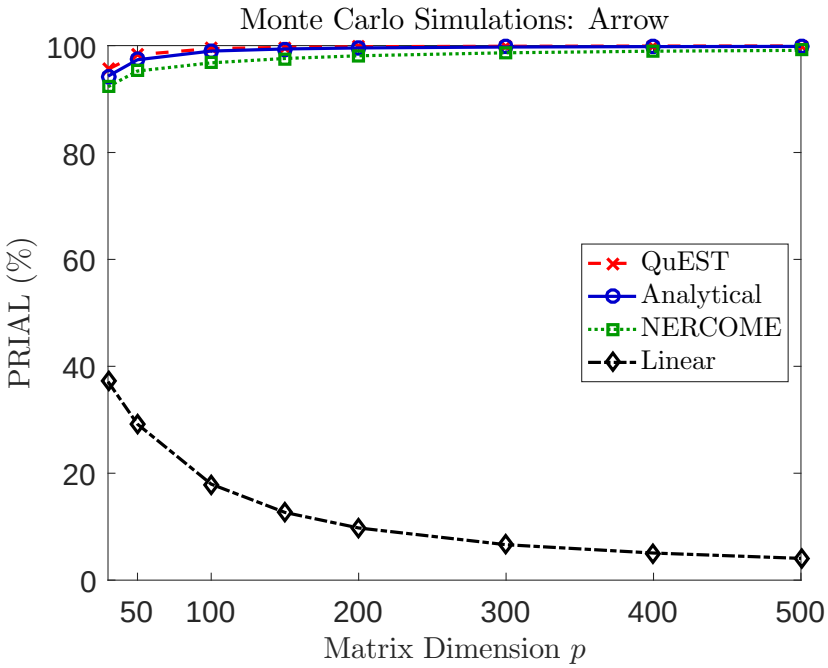

Figure 9: Evolution of the PRIAL as the matrix dimension, the top eigenvalue and the sample size all go to infinity together.

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

5.11 Summary

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

The three nonlinear shrinkage estimators perform very well across the board. Their performance levels are roughly similar to one another and of high standard. If anything, QuEST tends to be better than Analytical, which tends to be better than NERCOME, but the differences are relatively small, and there are exceptions. Between QuEST and Analytical there is hardly any difference at all.

The analytical nonlinear shrinkage formula 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 typically 1,000 times faster than the other nonlinear shrinkage methods, and is the only one that can handle ultra-large dimensions up to $10,000$ in reasonable time.

6 Robustness Checks for Kernel and Bandwidth

The goal of this section is to examine the extent to which the performance of the analytical nonlinear shrinkage estimator is sensitive to the choices of kernel and bandwidth.

6.1 Test Bench

To start with, we design a so-called "Test Bench" that spans most of the parameter variations reported above to comprehensively assess the performances of alternative estimation strategies. To this end, we scan 12 different combinations of (p, n) and 12 distributions of population eigenvalues, for a total of $12 \times 12 = 144$ scenarios.

For the dimension p, we use a low value of 50, a medium value of 100, and a high value of 200. For the concentration ratio c, we use four levels: 0.2, 0.4, 0.6, and 0.8. The condition number θ varies in the set $\{2, 10, 40\}$. Finally, we investigate four shapes for the population eigenvalues: uniform density; continuous bimodal Beta$(0.5, 0.5)$; discrete mass points conforming to the baseline scenario of Section 5.3; and the arrow model of Section 5.10. For each of the 144 scenarios, we run 100 Monte Carlo simulations.

6.2 Volatility Multiplier

Although the PRIAL is informative in controlled settings where we move one parameter at a time, it is not so easy to interpret across aggregated test bench results: Beating the sample covariance matrix decisively when it is much worse than FSOPT is more valuable than beating it when it is almost as good as FSOPT. To address this concern, we focus on a performance measure that has broad-spectrum practical interpretability. As shown by Engle et al. (2017, Section 4), and in line with the loss function of Section 2.2, the quantity

$$\mathrm{SD}(\widehat{\Sigma}_n, \Sigma_n) \coloneqq \frac{\sqrt{\mathsf{Tr}\big(\widehat{\Sigma}_n^{-1}\Sigma_n\widehat{\Sigma}_n^{-1}\big)/p}}{\mathsf{Tr}\big(\widehat{\Sigma}_n^{-1}\big)/p} \,. \tag{6.1}$$

is equivalent under large-dimensional asymptotics to the true (out-of-sample) standard deviation of a minimum-risk portfolio formed under arbitrary linear constraint, suitably normalized. In expression (6.1), Σ_n represents the population covariance matrix and $\widehat{\Sigma}_n$ a generic rotation-equivariant estimator of it, as described in Section 2. Proposition 2.1 shows that the finite-sample optimal (FSOPT) estimator S_n^* minimizes (6.1). Therefore, we report the "Volatility Multiplier"

$$\mathrm{VM}(\widehat{\Sigma}_n, \Sigma_n) \coloneqq \frac{\mathrm{SD}(\widehat{\Sigma}_n, \Sigma_n)}{\mathrm{SD}(S_n^*, \Sigma_n)} \,.$$

The volatility multiplier is the factor by which the out-of-sample volatility of a portfolio gets amplified due to less-than-perfect covariance matrix estimation. It is greater than or

equal to 1.00 in every single one of the 14,400 simulations by construction. In practice it is always strictly greater than 1.00 because hitting the lower bound of one would require oracle knowledge of the unobservable population covariance matrix Σ_n. Obviously, a lower VM is better, so the only question is how close *bona fide* estimators can get to the "speed of light" of 1.00. The main advantage of the VM is that it is easy to interpret economically in three complementary ways:

1. If the VM is 1.20 for example, it means that a portfolio manager who would have an annualized volatility of 10% (a common number) under ideal circumstances will experience instead 12% volatility due to covariance matrix estimation error, which constitutes an economically significant violation of her risk budget.
2. Conversely, if his investment acumen entitled her theoretically to an annualized Sharpe ratio of 1 (a fairly typical number also), covariance matrix estimation error would degrade his Sharpe ratio to $1/1.20 = 0.83$, once again an economically significant hit to a performance metric followed by investors.
3. Yet another way to see this is that the 'Student' t-statistic for her 5-year track record (a duration often looked at) moves from $\sqrt{5} = 2.24$, which is statistically significant at the usual level of 5%, down to $\sqrt{5}/1.20 = 1.86$, which is not (for a two-sided test).

There is nothing magical about a VM of 1.20: It was just chosen to illustrate how the variance multiplier directly impacts metrics that have intuitive meaning and importance. The VM for any candidate estimator is averaged over the 144 scenarios of the test bench in Section 6.1, and across 100 Monte Carlo simulations per scenario, for a total of 14,400 simulations per candidate estimator of the covariance matrix.

6.3 Initial Results

To inaugurate the test bench, Table 6 reports the VM for the competing estimators listed in Section 5.1.

Estimator	Sample	Linear	Analytical	QuEST	NERCOME	FSOPT
VM	1.45	1.07	1.01	1.01	1.01	1.00
Time (ms)	< 1	< 1	2	1,683	1,814	2

Table 6: Volatility amplification caused by covariance matrix estimation error.

Note that FSOPT has a volatility multiplier of 1.00 by construction. Although linear shrinkage goes a long way towards fixing the flaws of the sample covariance matrix, it leaves some room for another round of improvement. This is delivered by the three main

nonlinear shrinkage estimators: QuEST, NERCOME and Analytical enjoy near-perfect track record across the parameter space. Among them, only Analytical is as fast as FSOPT.

6.4 Numerical Isotonization

Stein (1986) launched the whole literature on optimal rotation-equivariant estimation of large-dimensional covariance matrices. His analytical formula was eerily prescient of ours, as detailed by Ledoit and Wolf (2018, p. 3810), even though he worked outside of Random Matrix Theory. Because Stein did not use nonparametric estimation of the Hilbert transform of the sample spectral density, his original estimator suffered much violation of eigenvalues ordering, which led him to post-processing through the *ad hoc* numerical method of isotonization. Rajaratnam and Vincenzi (2016) showed that this isotonization – a notoriously difficult procedure to analyze and justify theoretically – was the hidden key to the empirical success of Stein's (modified) estimator. So the first question mark is whether we should overlay isotonization to enforce perfect ordering of our nonlinearly shrunk eigenvalues?

To address this question, we compare the percentage of nearest-neighbor order violations for FSOPT with that for Analytical across the test bench. More formally, these percentages are defined as

$$\frac{1}{p-1}\sum_{i=1}^{p-1}\mathbb{1}_{\{d^*_{n,i}>d^*_{n,i+1}\}} \quad \text{and} \quad \frac{1}{p-1}\sum_{i=1}^{p-1}\mathbb{1}_{\{\widetilde{d}_{n,i}>\widetilde{d}_{n,i+1}\}}\ ,$$

for FSOPT and Analytical respectively. On average across the test bench, FSOPT has 53% violations whereas Analytical has only 34%. We also record the average magnitudes of the violations defined as

$$\frac{1}{a^*_n}\sum_{i=1}^{p-1}\frac{d^*_{n,i}-d^*_{n,i+1}}{\frac{1}{2}\left(d^*_{n,i}+d^*_{n,i+1}\right)}\mathbb{1}_{\{d^*_{n,i}>d^*_{n,i+1}\}} \quad \text{and} \quad \frac{1}{\widetilde{a}_n}\sum_{i=1}^{p-1}\frac{\widetilde{d}_{n,i}-\widetilde{d}_{n,i+1}}{\frac{1}{2}\left(\widetilde{d}_{n,i}+\widetilde{d}_{n,i+1}\right)}\mathbb{1}_{\{\widetilde{d}_{n,i}>\widetilde{d}_{n,i+1}\}}\ ,$$

for FSOPT and Analytical respectively, where

$$a^*_n := \sum_{i=1}^{p-1}\mathbb{1}_{\{d^*_{n,i}>d^*_{n,i+1}\}} \quad \text{and} \quad \widetilde{a}_n := \sum_{i=1}^{p-1}\mathbb{1}_{\{\widetilde{d}_{n,i}>\widetilde{d}_{n,i+1}\}}\ .$$

On average across the test bench, the average magnitude of violations is 4.4% for FSOPT versus 1.9% for Analytical. These findings indicate that there is no pressing necessity to restore order, as the finite-sample optimal estimator has even more disordered eigenvalues than our asymptotically optimal estimator.

Just to be sure, we also tried post-processing the analytical estimator through the so-called *Pool Adjacent Violators* (PAV) algorithm of Ayer et al. (1955), the most widespread numerical method of isotonization and obtained the same volatility multiplier of 1.01.

For all these reasons, we keep our estimator purely analytical and forgo isotonization.

6.5 Alternative Kernel Choices

In this section, we hold the bandwidth constant; therefore, all kernels are standardized to have unit variance like the Epanechnikov kernel of Equation (4.5) for the purpose of an apples-to-apples comparison.

6.5.1 Gaussian Kernel

First, we consider the well-known Gaussian kernel, also used by Jing et al. (2010):

$$\forall x \in \mathbb{R} \qquad \kappa^G(x) := \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \ .$$

This kernel violates Assumption 4.1 because its support is not compact, so from the point of view of theory we are in limbo. The expression for its Hilbert transform is not exactly trivial either. Gautschi and Waldvogel (2001) prove that

$$\forall x \in \mathbb{R} \qquad PV \int_{-\infty}^{+\infty} \frac{e^{-t^2}}{t-x} dt = -2\sqrt{\pi} D(x), \tag{6.2}$$

$$\text{where} \qquad D(x) := e^{-x^2} \int_0^x e^{t^2} dt$$

is the Dawson (1897) function. Dawson's integral is directly related to the *imaginary error function* erfi through $D(x) = \sqrt{\pi} e^{-x^2} \operatorname{erfi}(x)/2$, and erfi itself is simply the imaginary extension of the more famous real error function erf:

$$\forall x \in \mathbb{R} \qquad \operatorname{erfi}(x) := -i \operatorname{erf}(ix) \ , \quad \text{where } i := \sqrt{-1} \quad \text{and} \quad \operatorname{erf}(x) := \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \ .$$

The real error function is just a translation and rescaling of the standard normal cumulative distribution function Φ via the relation $\operatorname{erf}(x) = 2\Phi(x\sqrt{2}) - 1$.

The Dawson function is easy to invoke because it is prominent enough for standard numerical packages such as Matlab and Mathematica to include it pre-programmed. However, the underlying numerics are *very* slow. Rescaling expression (6.2) yields

$$\forall x \in \mathbb{R} \qquad \mathcal{H}_{\kappa^G}(x) = -\frac{\sqrt{2}}{\pi} D\left(\frac{x}{\sqrt{2}}\right)$$

6.5.2 Triangular Kernel

Another kernel is often encountered due to its simplicity: the triangular kernel

$$\forall x \in \mathbb{R} \qquad \kappa^T(x) := \frac{\sqrt{6} - |x|}{6} \mathbb{1}_{\{|x| \le \sqrt{6}\}} \ .$$

Poularikas (1998, Table 15.2) shows its Hilbert transform is

$$\forall x \notin \left\{-\sqrt{6}, 0, \sqrt{6}\right\} \qquad \mathcal{H}_{\kappa^T}(x) = \frac{1}{\sqrt{6}\pi} \left(\log \left| \frac{x - \sqrt{6}}{x + \sqrt{6}} \right| + \frac{x}{\sqrt{6}} \log \left| \frac{x^2}{x^2 - 6} \right| \right) \ ,$$

which is extended by continuity through $\mathcal{H}_{\kappa^T}\left(\pm\sqrt{6}\right) = \mp 2\log(2)/\left(\sqrt{6}\pi\right)$ and $\mathcal{H}_{\kappa^T}(0) = 0$.

6.5.3 Quartic Kernel

Its main attraction is that it is continuously differentiable even at the edges of the support:

$$\forall x \in \mathbb{R} \qquad \kappa^Q(x) := \frac{15}{16\sqrt{7}} \left(1 - \frac{x^2}{7}\right)^2 \mathbb{1}_{\{|x| \leq \sqrt{7}\}} \ .$$

Its Hilbert transform does not appear to have been computed in the literature. We derive it in the following proposition.

Proposition 6.1.

$$\forall x \in \mathbb{R} \quad \mathcal{H}_{\kappa^Q}(x) = \frac{15x^3 - 175x}{392\pi} + \begin{cases} \frac{15}{16\sqrt{7}\pi} \left(\frac{x^2}{7} - 1\right)^2 \log\left|\frac{\sqrt{7} - x}{\sqrt{7} + x}\right| & \text{if } |x| \neq \sqrt{7} \\ 0 & \text{if } |x| = \sqrt{7} \ . \end{cases}$$

6.5.4 Semicircular Kernel

The last alternative kernel is one that is not famous in the density estimation literature, but launched the whole literature on large-dimensional random matrices. It is known as the Wigner (1955) semicircular kernel. Wigner showed that the semicircle is the limiting sample spectral density of a class of real symmetric matrices under large-dimensional asymptotics. Given the direct lineage into our problem, we include the semicircular kernel for historical reasons. Its density and Hilbert transform have the relatively simple expressions already given in Table 1:

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

6.5.5 Comparison of Kernels

In order to build intuition, Figure 10 compares the alternative kernels and their Hilbert transforms with Epanechnikov. The semicircular kernel is omitted because it was already graphed in Figure 1.

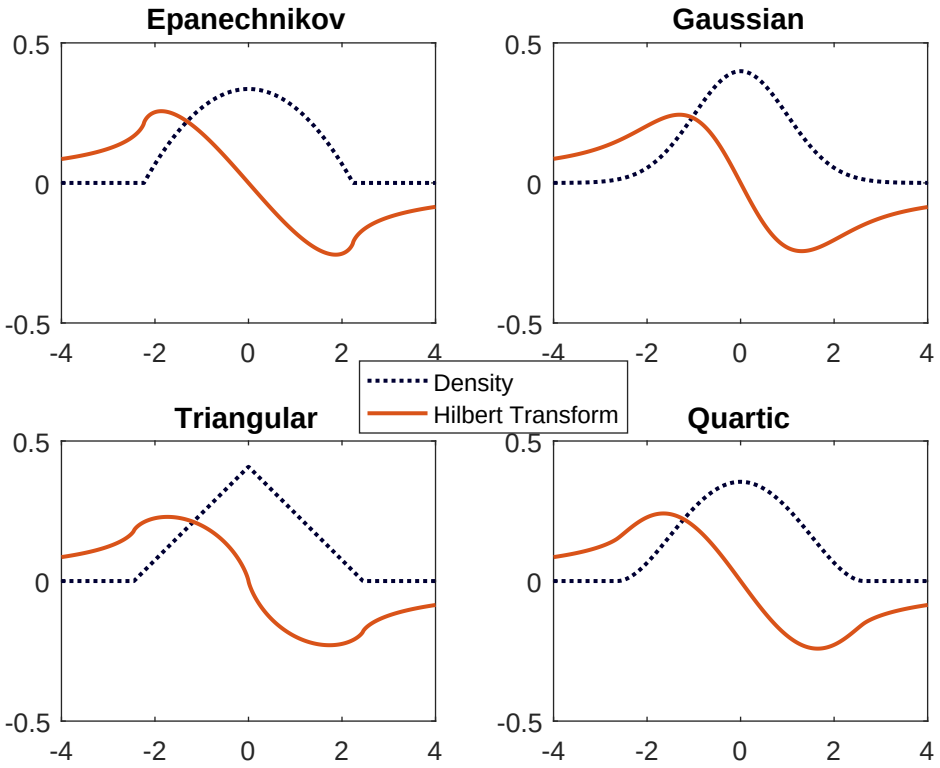

Figure 10: Comparison of unit-variance kernels and their Hilbert transforms.

Even though their formulas differ greatly, all Hilbert transforms behave similarly in that they signal which side is the closest mass: They are positive if more mass lies to the immediate right, are negative if more mass lies to the immediate left, and asymptote to zero away from mass. The results are presented in Table 7.

Kernel	Gaussian	Triangular	Quartic	Semicircular
VM	1.01	1.01	1.24	1.01
Time (ms)	10,066	2	3	2

Table 7: Robustness of performance to choice of kernel.

The triangular and semicircular kernels give the same accuracy and speed as Epanechnikov. The Gaussian kernel is extremely slow due to the evaluation of Dawson's integral, so this is an another reason to avoid it (in addition to its non-compact support).

An in-depth analysis of the woeful performance of the quartic kernel reveals that it is not attributable to a theoretical failure of the kernel itself, but to Matlab's inability to accurately compute $\mathcal{H}_{\kappa_Q}(x)$ for large x, due to the inherent limitations of double-precision arithmetic. When $x > 3,000$, Matlab has to subtract two quantities of order 10^{11}; and with 10^{-14} accuracy, the result is way off. The main culprit is the term $15x^3$ in

Proposition 6.1. Thus, we disqualify the quartic kernel for being guilty of numerical instability.

From these findings we conclude that the analytical nonlinear shrinkage formula of Section 4.7 is not sensitive to the choice of kernel, as long as we eschew the kernels with glaring numerical flaws: Gaussian (slow) and quartic (unstable). By default we keep Epanechnikov over triangular and semicircular because it has more established theoretical credentials and is more widely used, but the other two kernels seem to do just as well.

6.6 Locally Adaptive vs. Globally Uniform Bandwidth

Next, we scrutinize the contention of Section 4.2 that the intrinsic scale-equivariance of the problem demands that the bandwidth applied to sample eigenvalue $\lambda_{n,i}$ be proportionate to $\lambda_{n,i}$ itself ($i = 1, \ldots, p$). To this end, we modify the bandwidth $h_{n,i}$ of Section 4.6 to make it globally uniform, in the sense that it is scaled only to the *average* eigenvalue:

$$\forall i = 1, \ldots, p \qquad h_{n,i}^F := \left(\frac{1}{p} \sum_{j=1}^{p} \lambda_j \right) n^{-1/3} \ .$$

The second subscript i is redundant because the bandwidth is independent of i, but we preserve it for notational coherence with (4.9). For example, multiplying the data by two should result in a covariance matrix estimator multiplied by four: this universally accepted equivariance principle justifies multiplication by the average eigenvalue.

If we are going to consider globally uniform bandwidths, we might as well throw into the ring the often used data-driven rule-of-thumb of Silverman (1986):

$$\forall i = 1, \ldots, p \qquad h_{n,i}^S := \frac{2.345}{\sqrt{5}} n^{-1/5} \min \left(\widehat{\sigma}(\boldsymbol{\lambda}_n), \frac{\mathrm{IQR}(\boldsymbol{\lambda}_n)}{1.349} \right) \ ,$$

where $\widehat{\sigma}$ denotes the sample standard deviation and IQR denotes the inter-quartile range. Note that Cameron and Trivedi (2005, p. 304) do not divide by $\sqrt{5}$ because they rescale the Epanechnikov kernel to have support $[-1, 1]$. As required, multiplying the data by two makes $h_{n,i}^S$ four times as large.

Putting the globally uniform bandwidths $h_{n,i}^F$ and $h_{n,i}^S$ through the test bench yields volatility multipliers of 1.04 and 1.06, respectively. Unlike for linear shrinkage, they would eventually converge to 1.00 in the large-dimensional asymptotic limit, but for reasonable matrix dimensions $p \in \{50, 100, 200\}$, these performances lack luster, which vindicates our championing proportional bandwidth. Silverman's rule-of-thumb may be acceptable for generic density estimation; however, nonlinear shrinkage is *not generic* because the sample eigenvalues are not i.i.d. Our proportional bandwidth is a natural adaptation to the problem's idiosyncrasy.

6.7 Choices of Bandwidth Exponent and Scalar Multiplier

Finally, we revisit the decision made in Section 4.6 to set the global bandwidth h_n equal to $n^{-1/3}$. There was good *ex ante* justification to do so because 1/3 is the exponent championed by the only published paper on kernel estimation of the sample spectral density in large dimensions (Jing et al., 2010). Nonetheless, for the sake of completeness, we should also consider other exponents. Given that such exponents must be strictly less than 2/5 due to Theorem 4.1 and that 1/5 is a familiar exponent used in Silverman's rule of thumb, we consider the choices $\alpha \in \{0.2, 0.25, 0.30, 0.35\}$.

An additional degree of freedom is that it is possible to imagine putting a scalar K in front of $n^{-\alpha}$. We did not, which is the most neutral stance, but even this is implicitly a choice because it means we *de facto* chose $K = 1$. So we will consider variants $K \in \{0.5, 1, 2\}$. For the purpose of this section, $h_n := Kn^{-\alpha}$ and $h_{n,i} := K\lambda_i n^{-\alpha}$ $(i = 1, \ldots, p)$. The results are in presented Table 8.

VM	$K = 0.5$	$K = 1$	$K = 2$
$\alpha = 0.20$	1.01	1.01	1.04
$\alpha = 0.25$	1.01	1.01	1.02
$\alpha = 0.30$	1.02	1.01	1.01
$\alpha = 0.35$	1.02	1.01	1.01

Table 8: Effect of varying the exponent and scalar multiplier of the global bandwidth.

There certainly does not seem to be any case for introducing a scalar $K \neq 1$ as an artificial multiplier in front of $n^{-\alpha}$. For $K = 1$, the performance is essentially insensitive to the exponent, so by default we are happy to conserve $\alpha = 1/3$ unless and until new research supersedes the original paper of Jing et al. (2010).

6.8 Summary of Robustness Checks

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

7 Conclusion

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

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

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

The dimension of covariance matrices that can be handled successfully now is at least 10,000, one order of magnitude larger compared to the numerical nonlinear shrinkage methods of Ledoit and Wolf (2015) and Lam (2016), which is important in the age of Big Data. We trust that this feature will make *nonlinear* shrinkage even more attractive to applied researchers.

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 support of the limiting spectral distribution of large-dimensional random matrices. *Annals of Probability*, 26(1):316–345.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press, Cambridge.

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

Dawson, H. (1897). On the numerical value of $\int_0^h e^{x^2} dx$. *Proceedings of the London Mathematical Society*, 1(1):519–522.

Engle, R. F. and Colacito, R. (2006). Testing and valuing dynamic correlations for asset allocation. *Journal of Business & 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.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, 14(1):153–158.

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.

Gautschi, W. and Waldvogel, J. (2001). Computing the Hilbert transform of the generalized Laguerre and Hermite weight functions. *BIT Numerical Mathematics*, 41(3):490–503.

Girshick, M. A. (1939). On the sampling theory of roots of determinantal equations. *Annals of Mathematical Statistics*, 10(3):203–224.

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. *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*, 30(12):4349–4388.

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

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

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.

Poularikas, A. D., editor (1998). *The Handbook of Formulas and Tables for Signal Processing*. CRC Press, Boca Raton.

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

Ribes, A., Azaïs, J.-M., and Planton, S. (2009). Adaptation of the optimal fingerprint method for climate change detection using a well-conditioned covariance matrix estimate. *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* 1^{re} *Série*, 8(4):J1–J122.

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

A 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.

For notational simplicity, all the proofs assume that the support of F, denoted by $\mathsf{Supp}(F)$, is a finite interval $[a, b]$, where $0 < a < b < \infty$. At the cost of increased notational complexity, all the proofs also extend to the general case where $\mathsf{Supp}(F)$ 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 $0 < a_1 < b_1 < \cdots < a_\nu < b_\nu < \infty$.

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 3.1–3.3, 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 4.1 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 4.1 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 our proofs below, we will make use of the following result:

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

This result holds under our stated assumptions, which is seen as follows. The imaginary part has already been taken care of, as it is π times the derivative of the kernel density. As for the real part, it follows from the statement in Assumption 4.1 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, there exists $M < \infty$ such that*

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

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

Corollary A.1. *Under the assumptions of Lemma A.1, it then also holds that*

$$\sup_{x} \left| \breve{m}_{F_{c,H}}(x) \right| < \infty \ . \tag{A.3}$$

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

We work on an interval $[a', b']$ such that $0 < a' < a$ and $b < b' < +\infty$. Without loss of generality, in the following developments, we will work on a set of probability one on which F_n converges almost surely to F. We then take n large enough such that both $F_n(a') = 0$ and $F_n(b') = 1$, which can be done by the results of Bai and Silverstein (1998); in addition, n needs also to be large enough that both $F_{c_n,H_n}(a') = 0$ and $F_{c_n,H_n}(b') = 1$. First, we claim that

$$\sup_{x\in[a',b']} \left| \breve{m}_{\widetilde{F}_n}(x) - \int_{a'}^{b'} \frac{1}{th_n}\breve{m}_K\left(\frac{x-t}{th_n}\right) dF_{c_n,H_n}(t) \right| \longrightarrow 0 \qquad \text{(A.4)}$$

in probability. Indeed, from integration by parts,

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

where we have used Theorem 3 of Jing et al. (2010) in the last line, together with the fact that the multiplier between square brackets is $O(1)$ due to Lemma A.2 and result (A.1).

The next aim is to show that

$$
\int\frac{1}{th_n}\breve{m}_K\left(\frac{x-t}{th_n}\right)dF_{c_n,H_n}(t)-\int\frac{1}{th_n}\breve{m}_K\left(\frac{x-t}{th_n}\right)dF_{c,H}(t)\longrightarrow 0 \tag{A.5}
$$

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,

$$
\int\frac{1}{1+uh_n}\breve{m}_K(u)\left[F'_{c_n,H_n}\left(\frac{x_n}{1+uh_n}\right)-F'_{c,H}\left(\frac{x_n}{1+uh_n}\right)\right]du\longrightarrow 0\ . \tag{A.6}
$$

From Theorem 1.1 of Silverstein and Choi (1995), $F'_{c,H}$ is uniformly bounded on $\mathsf{Supp}(F)$. Therefore, (A.6) follows from the dominated convergence theorem, Lemma A.1, Corollary 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

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

Therefore,

$$\begin{aligned}
&\sup_{x\in[a,b]} \left| \int \frac{1}{th_n}\mathsf{Re}\left[\breve{m}_K\left(\frac{x-t}{th_n}\right)\right] dF_{c,H}(t) - \mathsf{Re}[\breve{m}_{F_{c,H}}(x)] \right| \\
&\quad = \sup_{x\in[a,b]} \left| \int_{-R}^{R} \frac{1}{1+vh_n} k(v)\mathsf{Re}\left[\breve{m}_{F_{c,H}}\left(\frac{x}{1+vh_n}\right)\right] dv - \mathsf{Re}[\breve{m}_{F_{c,H}}(x)] \right| \\
&\quad \leq \sup_{x\in[a,b]} \left| \int_{-R}^{R} \frac{1}{1+vh_n} k(v) \left\{ \mathsf{Re}\left[\breve{m}_{F_{c,H}}\left(\frac{x}{1+vh_n}\right)\right] - \mathsf{Re}[\breve{m}_{F_{c,H}}(x)] \right\} dv \right| \\
&\qquad + \sup_{x\in[a,b]} \left| \mathsf{Re}[\breve{m}_F(x)] \right| \times \left| 1 - \int_{-R}^{R} \frac{1}{1+vh_n} k(v) dv \right| \ .
\end{aligned} \tag{A.7}$$

Note that, by the dominated convergence theorem,

$$\lim_{n\to\infty} \int_{-R}^{R} \frac{1}{1+vh_n} k(v)dv = \int_{-R}^{R} k(v)dv = 1 \ . \tag{A.8}$$

Therefore, the second term in (A.7) is $o(1)$. And for the first term it holds that

$$\begin{aligned}
&\sup_{x\in[a,b]} \left| \int_{-R}^{R} \frac{1}{1+vh_n} k(v) \left\{ \mathsf{Re}\left[\breve{m}_{F_{c,H}}\left(\frac{x}{1+vh_n}\right)\right] - \mathsf{Re}[\breve{m}_{F_{c,H}}(x)] \right\} dv \right| \\
&\quad \leq \sup_{x\in[a,b]} \int_{-R}^{R} \frac{1}{1+vh_n} k(v) \left| \mathsf{Re}\left[\breve{m}_{F_{c,H}}\left(\frac{x}{1+vh_n}\right)\right] - \mathsf{Re}[\breve{m}_{F_{c,H}}(x)] \right| dv \\
&\quad \leq \sup_{\substack{x,y\in\left[a-\frac{Rh_n}{1-Rh_n}, b+\frac{Rh_n}{1-Rh_n}\right] \\ |x-y|\leq \frac{Rh_n}{1-Rh_n}}} \left| \mathsf{Re}\left[\breve{m}_{F_{c,H}}(y)\right] - \mathsf{Re}[\breve{m}_{F_{c,H}}(x)] \right| \times \left| \int_{-R}^{R} \frac{1}{1+vh_n} k(v)dv \right| \\
&\quad \leq \sup_{\substack{x,y\in\left[\frac{a}{2}, b+\frac{a}{2}\right] \\ |x-y|\leq \frac{Rh_n}{1-Rh_n}}} \left| \mathsf{Re}\left[\breve{m}_{F_{c,H}}(y)\right] - \mathsf{Re}[\breve{m}_{F_{c,H}}(x)] \right| \times \left| \int_{-R}^{R} \frac{1}{1+vh_n} k(v)dv \right| \text{for large } n.
\end{aligned} \tag{A.9}$$

By the Heine-Cantor theorem, the first term of expression (A.9) converges to zero and by (A.8) the second expression of (A.9) converges to one. This guarantees that the bound (A.9) 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

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

A closely related statement, namely

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

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.10), 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.9) and (A.10) together imply

$$\sup_{x\in[a,b]}\left|\int \frac{1}{th_n}\breve{m}_K\left(\frac{x-t}{th_n}\right)dF_{c,H}(t)-\breve{m}_{F_{c,H}}(x)\right|\longrightarrow 0\ . \tag{A.12}$$

Results (A.4), (A.5), and (A.12) 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 4.1

We rescale the kernel to streamline calculations:

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

Elementary algebra enables us to verify that

$$\forall t\neq x\qquad \frac{1-t^2}{t-x}=-(t-x)-2x+\frac{1-x^2}{t-x}\ .$$

Therefore, if $x\notin[t_1,t_2]$,

$$\int_{t_1}^{t_2}\frac{1-t^2}{t-x}dt=\int_{t_1}^{t_2}\left[-(t-x)-2x+\frac{1-x^2}{t-x}\right]dt=\int_{t_1-x}^{t_2-x}\left[-u-2x+\frac{1-x^2}{u}\right]du\ ,$$

where we applied the change of variable $u = t - x$. Outside the support of the rescaled kernel $\underline{\kappa}^E$, the Cauchy Principal Value is a standard integral:

$$\forall x \notin [-1,1] \qquad \mathcal{H}_{\underline{\kappa}^E}(x) = \frac{1}{\pi}\int_{-1}^{1} \frac{3}{4}\frac{1-t^2}{1-x}dt = \frac{3}{4\pi}\left[-\frac{u^2}{2} - 2xu + (1-x^2)\log|u|\right]_{-1-x}^{1-x}$$
$$= \frac{3}{4\pi}\left[-2x + (1-x^2)\log\left|\frac{1-x}{1+x}\right|\right] .$$

Computations in the interior of the support are more complicated:

$$\forall x \in (-1,1) \qquad \mathcal{H}_{\underline{\kappa}^E}(x) = -\frac{3x}{2\pi} + \frac{3(1-x^2)}{4\pi}\lim_{\varepsilon\searrow 0}\left\{\left[\log|u|\right]_{-1-x}^{-\varepsilon} + \left[\log|u|\right]_{\varepsilon}^{1-x}\right\}$$
$$= -\frac{3x}{2\pi} + \frac{3(1-x^2)}{4\pi}\lim_{\varepsilon\searrow 0}\left\{\log\left|\frac{\varepsilon}{1+x}\right| + \log\left|\frac{1-x}{\varepsilon}\right|\right\}$$
$$= \frac{3}{4\pi}\left[-2x + (1-x^2)\log\left|\frac{1-x}{1+x}\right|\right] ,$$

but they come up with the same formula in the end. Values at the edge of the support are obtained by continuity as $\mathcal{H}_{\underline{\kappa}^E}(\pm 1) = \mp 3/(2\pi)$. Proposition 4.1 then follows by applying formula (4) of Erdélyi (1954, Section 15.1):

$$\forall x \in \mathbb{R} \qquad \mathcal{H}_{\kappa^E}(x) = \frac{1}{\sqrt{5}}\mathcal{H}_{\underline{\kappa}^E}\left(\frac{x}{\sqrt{5}}\right) . \blacksquare$$

A.7 Proof of Proposition 4.2

Simply from its definition, it is obvious that the Epanechnikov kernel (4.5) 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 follows from Proposition 4.1. It is continuous because

$$\lim_{x\to\pm\sqrt{5}}\left(1-\frac{x^2}{5}\right)\log\left|\frac{\sqrt{5}-x}{\sqrt{5}+x}\right| = 0 .$$

The Epanechnikov kernel (4.5) is a function of bounded variation because it is increasing on $[-\sqrt{5},0]$, decreasing on $[0,\sqrt{5}]$, and constant everywhere else. As for proving that its Hilbert transform (4.6) is a function of bounded variation, first notice that by symmetry, we need only consider the positive half of the real line. For every ε in $(0,\sqrt{5})$, $\mathcal{H}_{\kappa^E}$ is of bounded variation on $[0,\sqrt{5}-\varepsilon]$ because it is continuously differentiable on a compact interval. Furthermore, $\mathcal{H}_{\kappa^E}$ is of bounded variation on $[\sqrt{5}+\varepsilon,+\infty)$ because it is nondecreasing on this interval (this is because the support of κ^E is to the left of the interval), so it is bounded below by $-3\sqrt{5}/(10\pi)$ and above by zero. All that is left is to elucidate is the behavior in a neighborhood of $\sqrt{5}$. We can ignore the linear part $-3x/(10\pi)$ because it is evidently of bounded variation, so we are left with the function

$$\forall x \in \left[\sqrt{5}-\varepsilon,\sqrt{5}\right) \cup \left(\sqrt{5},\sqrt{5}+\varepsilon\right] \qquad g^E(x) := \left(1-\frac{x^2}{5}\right)\log\left|\frac{\sqrt{5}-x}{\sqrt{5}+x}\right| ,$$

extended by continuity through $g^E(\sqrt{5}) = 0$. It is also of bounded variation for sufficiently small $\varepsilon > 0$. This because its derivative

$$\frac{dg^E}{dx}(x) = -\frac{2x}{5}\log\left|\frac{\sqrt{5}-x}{\sqrt{5}+x}\right| - \frac{2}{\sqrt{5}}\ ,$$

tends to $+\infty$ as x tends to $\sqrt{5}$ from either side and is therefore strictly positive on $\left[\sqrt{5}-\varepsilon, \sqrt{5}\right) \cup \left(\sqrt{5}, \sqrt{5}+\varepsilon\right]$ for sufficiently small ε. This fact, together with the continuity of the function g^E itself, implies that g^E is monotonically increasing, and thus of bounded variation, on $\left[\sqrt{5}-\varepsilon, \sqrt{5}+\varepsilon\right]$ for sufficiently small ε. ■

A.8 Proof of Proposition 6.1

The rescaled function

$$\forall x \in \mathbb{R} \qquad \underline{\kappa}^Q(x) := \frac{16\sqrt{7}}{15}\,\kappa^Q\left(\sqrt{7}x\right) = \left(1-x^2\right)^2 \mathbb{1}_{\{|x|\leq 1\}}$$

is easier to analyze. Tedious algebra confirms that for all $t, x \in \mathbb{R}$,

$$\frac{(1-t^2)^2}{t-x} = (t-x)^3 + 4x(t-x)^2 + (6x^2-2)(t-x) + 4x(x^2-1) + \frac{(x^2-1)^2}{t-x}\ .$$

The change of variable $u = t - x$ yields for any $x \notin [t_1, t_2]$

$$\begin{aligned}\int_{t_1}^{t_2} \frac{(1-t^2)^2}{t-x}dt &= \int_{t_1-x}^{t_2-x}\left[u^3 + 4xu^2 + (6x^2-2)u + 4x(x^2-1) + \frac{(x^2-1)^2}{u}\right]du \\ &= \left[\frac{u^4}{4} + \frac{4xu^3}{3} + (3x^2-1)u^2 + 4x(x^2-1)u + \left(x^2-1\right)^2\log|u|\right]_{t_1-x}^{t_2-x}\ .\end{aligned}$$

Outside the support of $\underline{\kappa}^Q$, the Principal Value is a standard integral: for all $x \notin [-1, 1]$

$$\begin{aligned}\mathcal{H}_{\underline{\kappa}^Q}(x) &= \frac{1}{\pi}\int_{-1}^{1}\frac{(1-t^2)^2}{1-x}dt \\ &= \frac{1}{\pi}\left[\frac{u^4}{4} + \frac{4xu^3}{3} + (3x^2-1)u^2 + 4x(x^2-1)u + \left(x^2-1\right)^2\log|u|\right]_{-1-x}^{1-x} \\ &= \frac{1}{\pi}\left[2x^3 - \frac{10}{3}x + (x^2-1)^2\log\left|\frac{1-x}{1+x}\right|\right]\ .\end{aligned}$$

Computations in the interior of the support are more complicated:

$$\begin{aligned}\forall x \in (-1,1) \quad \mathcal{H}_{\underline{\kappa}^Q}(x) &= \frac{1}{\pi}\left[2x^3 - \frac{10}{3}x + (x^2-1)^2\lim_{\varepsilon\searrow 0}\left\{\left[\log|u|\right]_{-1-x}^{-\varepsilon} + \left[\log|u|\right]_{\varepsilon}^{1-x}\right\}\right] \\ &= \frac{1}{\pi}\left[2x^3 - \frac{10}{3}x + (x^2-1)^2\lim_{\varepsilon\searrow 0}\left\{\log\left|\frac{\varepsilon}{1+x}\right| + \log\left|\frac{1-x}{\varepsilon}\right|\right\}\right] \\ &= \frac{1}{\pi}\left[2x^3 - \frac{10}{3}x + (x^2-1)^2\log\left|\frac{1-x}{1+x}\right|\right]\ ,\end{aligned}$$

but they end up with the same formula. Values at the edge of the support are obtained by continuity as $\mathcal{H}_{\underline{\kappa}^E}(\pm 1) = \mp 4/(3\pi)$. Proposition 6.1 then follows from formula (4) of Erdélyi (1954, Section 15.1):

$$\begin{aligned}\forall x \in \mathbb{R} \quad \mathcal{H}_{\kappa^Q}(x) &= \frac{15}{16\sqrt{7}}\,\mathcal{H}_{\underline{\kappa}^Q}\left(\frac{x}{\sqrt{7}}\right) \\ &= \frac{15x^3 - 175x}{392\pi} + \begin{cases} \dfrac{15}{16\sqrt{7}\pi}\left(\dfrac{x^2}{7} - 1\right)^2 \log\left|\dfrac{\sqrt{7}-x}{\sqrt{7}+x}\right| & \text{if } |x| \neq \sqrt{7} \\ 0 & \text{if } |x| = \sqrt{7}\ . \ \blacksquare \end{cases}\end{aligned}$$

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 5.5, 5.6, and 5.9 — 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 the loss function. The many applied researchers 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, defined in a manner analogous to Equation (5.1) as

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

B.1 Concentration Ratio

First we revisit the results of Section 5.5, 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 11 below.

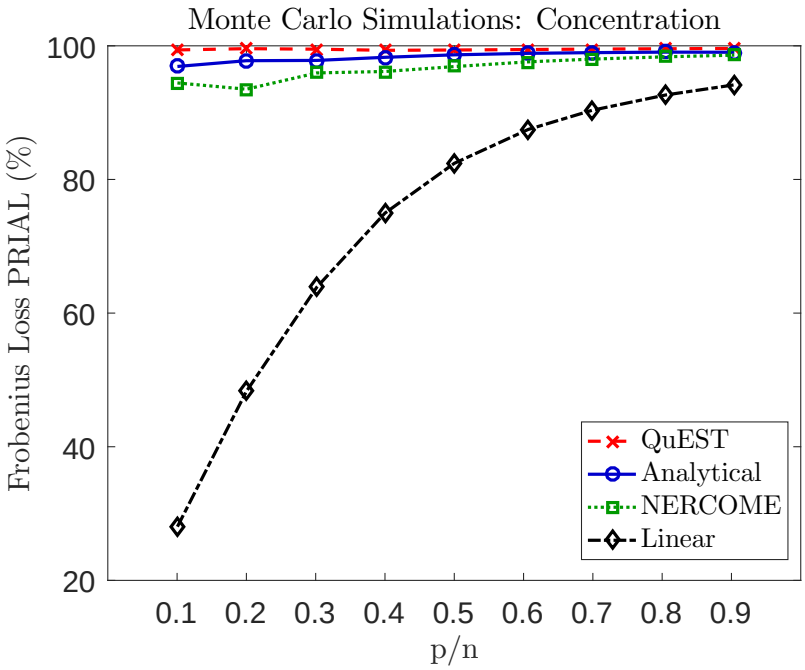

Figure 11: Evolution of the Frobenius PRIAL of various estimators as a function of the ratio of matrix dimension to sample size.

Linear shrinkage is now comfortably above the sample covariance matrix. This confirms that any underperformance observed in Section 5.5 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 5.6, 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 12 below.

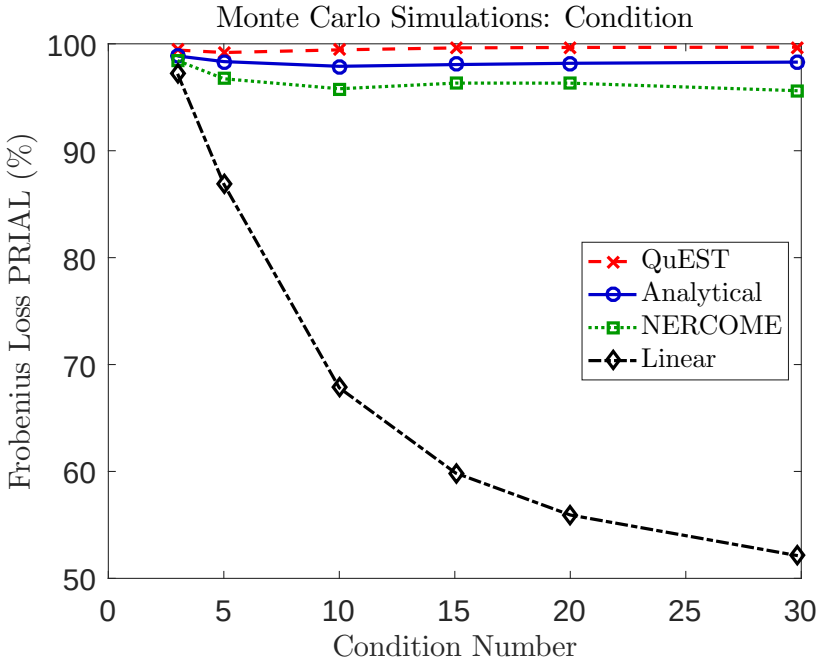

Figure 12: Evolution of the Frobenius PRIAL of various estimators as a function of the condition number of the population covariance matrix.

Linear shrinkage is also comfortably above the sample covariance matrix. Any underperformance observed in Section 5.5 is solely attributable to the loss function.

B.3 Fixed-Dimensional Asymptotics

Third and last, we revisit the results of Section 5.9, 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 13 below.

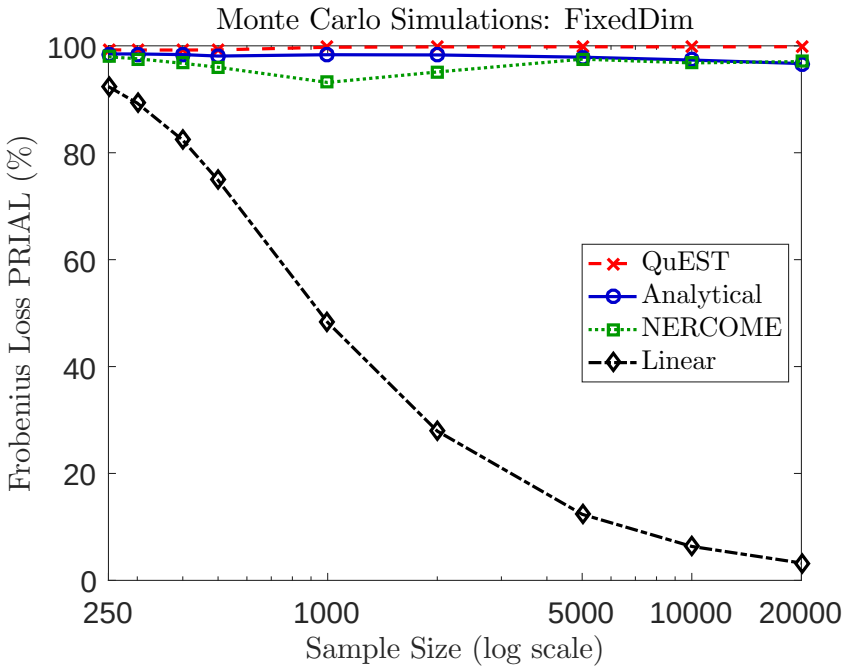

Figure 13: Evolution of the Frobenius PRIAL of various estimators as the sample size grows towards infinity, while the matrix dimension remains fixed.

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 3.1. The oracle nonlinear shrinkage function defined in Equation (3.3) can be rewritten in terms of these new objects as

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

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 $\underline{f}$ by

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

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

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

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

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

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,

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

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

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

As before, we operationalize these formulas with the Epanechnikov (1969) kernel $\kappa^E(x)$ and the proportional bandwidth $h_{n,i} = \lambda_{n,i} h_n$: From this we deduce for all $i = p-n+1, \ldots, p$,

$$\widetilde{\underline{f}}_n(\lambda_{n,i}) = \frac{1}{n} \sum_{j=p-n+1}^{p} \frac{3}{4\sqrt{5}\lambda_{n,j} h_n} \left[1 - \frac{1}{5} \left(\frac{\lambda_{n,i} - \lambda_{n,j}}{\lambda_{n,j} h_n} \right)^2 \right]^+ \tag{C.6}$$

$$\mathcal{H}_{\widetilde{\underline{f}}_n}(\lambda_{n,i}) = \frac{1}{n} \sum_{j=p-n+1}^{p} \left\{ -\frac{3(\lambda_{n,i} - \lambda_{n,j})}{10\pi\lambda_{n,j}^2 h_n^2} + \frac{3}{4\sqrt{5}\pi\lambda_{n,j} h_n} \left[1 - \frac{1}{5} \left(\frac{\lambda_{n,i} - \lambda_{n,j}}{\lambda_{n,j} h_n} \right)^2 \right] \times \log \left| \frac{\sqrt{5}\lambda_{n,j} h_n - \lambda_{n,i} + \lambda_{n,j}}{\sqrt{5}\lambda_{n,j} h_n + \lambda_{n,i} - \lambda_{n,j}} \right| \right\} . \tag{C.7}$$

$$\mathcal{H}_{\widetilde{\underline{f}}_n}(0) = \left[\frac{3}{10h_n^2} + \frac{3}{4\sqrt{5}h_n} \left(1 - \frac{1}{5h_n^2} \right) \log \left(\frac{1+\sqrt{5}h_n}{1-\sqrt{5}h_n} \right) \right] \times \underbrace{\frac{1}{\pi n} \sum_{j=p-n+1}^{p} \frac{1}{\lambda_{n,j}}}_{\mathcal{H}_{\underline{F}'_n}(0)} . \tag{C.8}$$

Note that, as $n \to \infty$, the bracketed expression in front of $\mathcal{H}_{\underline{F}'_n}(0)$ in (C.8) converges 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 $\widetilde{\boldsymbol{d}}_n$ and recompose them with the sample eigenvectors to compute the covariance matrix estimator $\widetilde{S}_n := \sum_{i=1}^{p} \widetilde{d}_{n,i} \cdot u_{n,i} u'_{n,i}$.

Doing so enables us to run a counterpart of the Monte Carlo simulations in Section 5.5 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 14. Given that the minimum variance loss $\mathcal{L}_n^{\text{MV}}$ 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 $\mathcal{L}_n^{\text{FR}}$ 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.

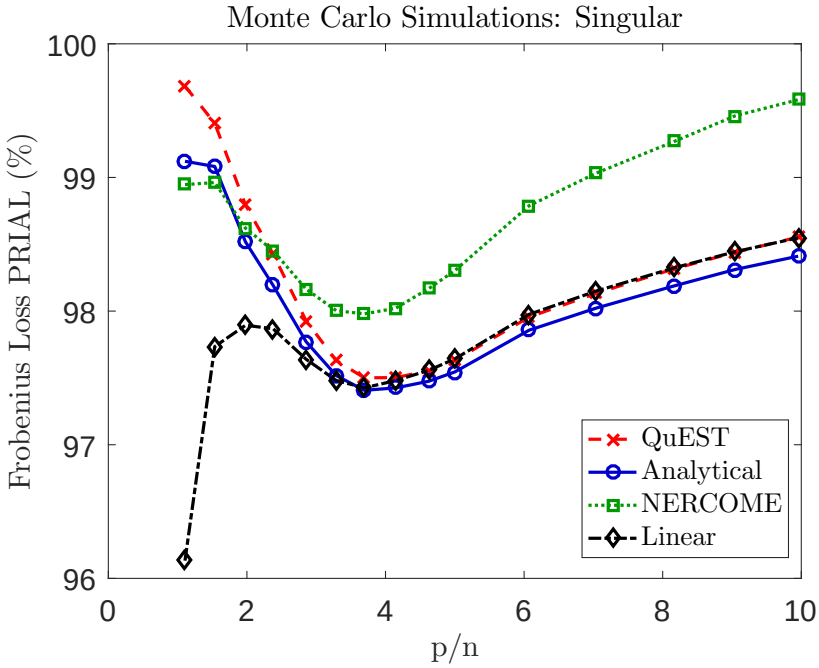

Figure 14: Evolution of the Frobenius PRIAL of various estimators when the matrix dimension exceeds the sample size.

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 now not so large anymore.

D Programming Code

The Matlab function that computes the nonlinear shrinkage estimator of the covariance matrix based on our new methodology has only about 20 to 30 lines of actual Matlab code, which makes for easy debugging and customization.

```
function sigmatilde=analytical_shrinkage(X)
% extract sample eigenvalues sorted in ascending order and eigenvectors
[n,p]=size(X); % important: sample size n must be >= 12
sample=(X'*X)./n;
[u,lambda]=eig(sample,'vector');
[lambda,isort]=sort(lambda);
u=u(:,isort);
% compute analytical nonlinear shrinkage kernel formula
lambda=lambda(max(1,p-n+1):p);
L=repmat(lambda,[1 min(p,n)]);
h=n^(-1/3);                                          % Equation (4.9)
H=h*L';
x=(L-L')./H;
ftilde=(3/4/sqrt(5))*mean(max(1-x.^2./5,0)./H,2);    % Equation (4.7)
Hftemp=(-3/10/pi)*x+(3/4/sqrt(5)/pi)*(1-x.^2./5) ...
   .*log(abs((sqrt(5)-x)./(sqrt(5)+x)));             % Equation (4.8)
Hftemp(abs(x)==sqrt(5))=(-3/10/pi)*x(abs(x)==sqrt(5));
Hftilde=mean(Hftemp./H,2);
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/pi)*(3/10/h^2+3/4/sqrt(5)/h*(1-1/5/h^2) ...
      *log((1+sqrt(5)*h)/(1-sqrt(5)*h)))*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
sigmatilde=u*diag(dtilde)*u';                        % Equation (4.4)
```

The analytical nonlinear shrinkage 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 `sigmatilde`.