Direct Nonlinear Shrinkage Estimation of Large-Dimensional Covariance Matrices

Olivier Ledoit$^1$ and Michael Wolf$^1$

$^1$Department of Economics
University of Zurich
Outline

1. The Problem
2. Finite Samples
3. Large-Dimensional Asymptotics
4. Kernel Estimation
5. Monte Carlo Study
6. Conclusion
Outline

1. The Problem
2. Finite Samples
3. Large-Dimensional Asymptotics
4. Kernel Estimation
5. Monte Carlo Study
6. Conclusion
Problem:

- We want to estimate a $p$-dimensional covariance matrix based on an i.i.d. sample of size $n$
- The classic estimator is the sample covariance matrix $S_n$
- However, this estimator is ill-conditioned when $p$ is of the same magnitude as $n$, and tends to perform poorly
Problem:

- We want to estimate a $p$-dimensional covariance matrix based on an i.i.d. sample of size $n$
- The classic estimator is the sample covariance matrix $S_n$
- However, this estimator is ill-conditioned when $p$ is of the same magnitude as $n$, and tends to perform poorly

Note:

- One of the most important problems in multivariate statistics
- Applications are plentiful
Previous Approaches

(1) **Incorporate additional knowledge** in the estimation process:

- Rely on a sparsity, such as Bickel and Levina (2008, AoS)
- Rely on a graph model, such as Rajaratnam et al. (2008, AoS)
- Rely on a factor structure, such as Fan et al. (2013, JRSS-B)
Previous Approaches

(1) **Incorporate additional knowledge** in the estimation process:
- Rely on a sparsity, such as Bickel and Levina (2008, AoS)
- Rely on a graph model, such as Rajaratnam et al. (2008, AoS)
- Rely on a factor structure, such as Fan et al. (2013, JRSS-B)

(2) **Linear shrinkage:**
- Consider estimators of the form:

\[ \delta \cdot \bar{s}_{n}^2 \cdot I_p + (1 - \delta) \cdot S_n \]

where \( \bar{s}_{n}^2 \) is the grand mean of the sample variances \( s_{n,i}^2 \).  
- Ledoit and Wolf (2004, JMVA) derive asymptotically optimal
  *bona fide* shrinkage intensity \( \delta \) under the Frobenius loss.
Linear Shrinkage

Immediate interpretation:

- Shrink the elements of $S_n$ to the elements of $\bar{s}_n^2 \cdot I_p$ with common intensity $\delta$
Linear Shrinkage

Immediate interpretation:

- Shrink the elements of $S_n$ to the elements of $\bar{s}_n^2 \cdot I_p$
  with common intensity $\delta$

Alternative interpretation:

- Decompose the sample covariance matrix into eigenvalues and eigenvectors: $\{ (\lambda_{n,1}, \ldots, \lambda_{n,p}); (u_{n,1}, \ldots, u_{n,p}) \}$
- Keep the sample eigenvectors
- Shrink the sample eigenvalues $\lambda_{n,i}$ to their grand mean $\bar{\lambda}_n$
  with common intensity $\delta$:

$$\lambda_{n,i}^{\text{shrunk}} := \delta \bar{\lambda}_n + (1 - \delta) \lambda_{n,i}$$
Linear Shrinkage

Immediate interpretation:
- Shrink the elements of $S_n$ to the elements of $\bar{s}_n^2 \cdot I_p$
  with common intensity $\delta$

Alternative interpretation:
- Decompose the sample covariance matrix into eigenvalues and eigenvectors: $\{(\lambda_{n,1}, \ldots, \lambda_{n,p}); (u_{n,1}, \ldots, u_{n,p})\}$
- Keep the sample eigenvectors
- Shrink the sample eigenvalues $\lambda_{n,i}$ to their grand mean $\bar{\lambda}_n$
  with common intensity $\delta$:

$$\lambda_{n,i}^{\text{shrunken}} := \delta \bar{\lambda}_n + (1 - \delta) \lambda_{n,i}$$

In particular:
- The shrunken eigenvalues are obtained by applying a linear transformation to the sample eigenvalues
Nonlinear Shrinkage

More general approach:

- Decompose the sample covariance matrix into eigenvalues and eigenvectors: \(\{ (\lambda_{n,1}, \ldots, \lambda_{n,p}); (u_{n,1}, \ldots, u_{n,p}) \}\)
- Keep the sample eigenvectors
- Shrink the sample eigenvalues \(\lambda_{n,i}\) to their grand mean \(\bar{\lambda}_n\), but at distinct intensities (even allowed to be negative)
Nonlinear Shrinkage

More general approach:

- Decompose the sample covariance matrix into eigenvalues and eigenvectors: \( \{(\lambda_{n,1}, \ldots, \lambda_{n,p}); (u_{n,1}, \ldots, u_{n,p})\} \)
- Keep the sample eigenvectors
- Shrink the sample eigenvalues \( \lambda_{n,i} \) to their grand mean \( \bar{\lambda}_n \), but at distinct intensities (even allowed to be negative)

In particular:

- The shrunken eigenvalues are obtained by applying a nonlinear transformation to the sample eigenvalues
Nonlinear Shrinkage

More general approach:

- Decompose the sample covariance matrix into eigenvalues and eigenvectors: \((\lambda_{n,1}, \ldots, \lambda_{n,p}); (u_{n,1}, \ldots, u_{n,p})\)
- Keep the sample eigenvectors
- Shrink the sample eigenvalues \(\lambda_{n,i}\) to their grand mean \(\bar{\lambda}_n\), but at distinct intensities (even allowed to be negative)

In particular:

- The shrunken eigenvalues are obtained by applying a nonlinear transformation to the sample eigenvalues

Doing so should yield even better results, if done right.
Outline

1 The Problem
2 Finite Samples
3 Large-Dimensional Asymptotics
4 Kernel Estimation
5 Monte Carlo Study
6 Conclusion
Reasonable Restriction

Rotation-Equivariant Estimators

- $Y_n$ are the observed data, an $n \times p$ matrix
- $W$ is a $p \times p$ orthogonal matrix
- $\hat{\Sigma}_n := \hat{\Sigma}_n(Y_n)$ is an estimator of $\Sigma_n$
- It is rotation-equivariant if $\hat{\Sigma}_n(Y_nW) = W'\hat{\Sigma}_n(Y_n)W$

Without specific knowledge about $\Sigma_n$, it is reasonable to restrict attention to this class of estimators.
Reasonable Restriction

Rotation-Equivariant Estimators

- $Y_n$ are the observed data, an $n \times p$ matrix
- $W$ is a $p \times p$ orthogonal matrix
- $\hat{\Sigma}_n := \hat{\Sigma}_n(Y_n)$ is an estimator of $\Sigma_n$
- It is rotation-equivariant if $\hat{\Sigma}_n(Y_nW) = W'\hat{\Sigma}_n(Y_n)W$

Without specific knowledge about $\Sigma_n$, it is reasonable to restrict attention to this class of estimators.

We use the following class of rotation-equivariant estimators going back to Stein (1975, 1986):

$$\hat{\Sigma}_n := U_n\hat{\Delta}_n U_n'$$

where

$$\hat{\Delta}_n := \text{Diag}(\hat{\delta}_{n,1}, \ldots, \hat{\delta}_{n,p})$$
is diagonal
Finite-Sample Optimality

Starting objective:

- Find the matrix in this class that is closest to \( \Sigma_n \)
- Distance is measured by the minimum-variance loss

\[
\mathcal{L}_n^{MV}(\tilde{\Sigma}_n, \Sigma_n) := \frac{\text{Tr}(\tilde{\Sigma}_n^{-1} \Sigma_n \tilde{\Sigma}_n^{-1})/p}{[\text{Tr}(\tilde{\Sigma}_n^{-1})/p]^2} - \frac{1}{\text{Tr}(\Sigma_n^{-1})/p}
\]
Finite-Sample Optimality

Starting objective:

- Find the matrix in this class that is closest to $\Sigma_n$
- Distance is measured by the minimum-variance loss

$$\mathcal{L}_n^{MV}(\hat{\Sigma}_n, \Sigma_n) := \frac{\text{Tr}(\hat{\Sigma}_n^{-1} \Sigma_n \hat{\Sigma}_n^{-1})}{p} - \frac{1}{\left[ \text{Tr}(\hat{\Sigma}_n^{-1})/p \right]^2} - \frac{\text{Tr}(\Sigma_n^{-1})}{p}$$

Minimization problem:

$$\min_{\hat{\Delta}_n} \mathcal{L}_n^{MV}(U_n \hat{\Delta}_n U'_n, \Sigma_n)$$
Finite-Sample Optimality

Starting objective:
- Find the matrix in this class that is closest to $\Sigma_n$
- Distance is measured by the minimum-variance loss

$$\mathcal{L}_{n}^{MV}(\Sigma_n, \Sigma_n) := \frac{\text{Tr}(\Sigma_n^{-1} \Sigma_n \Sigma_n^{-1})/p}{\left[ \text{Tr}(\Sigma_n^{-1})/p \right]^2} - \frac{1}{\text{Tr}(\Sigma_n^{-1})/p}$$

Minimization problem:
$$\min_{\hat{\Delta}_n} \mathcal{L}_{n}^{MV}(U_n \hat{\Delta}_n U'_n, \Sigma_n)$$

Solution:
$$\Delta^*_n := \text{Diag}(\delta^*_n, 1, \ldots, \delta^*_n, p) \quad \text{where} \quad \delta^*_n := u'_{n,i} \Sigma_n u_{n,i}$$
Finite-Sample Optimality

Starting objective:

- Find the matrix in this class that is closest to $\Sigma_n$
- Distance is measured by the minimum-variance loss

$$\mathcal{L}_{n}^{\text{MV}}(\hat{\Sigma}_n, \Sigma_n) := \frac{\text{Tr}(\hat{\Sigma}_n^{-1}\Sigma_n\hat{\Sigma}_n^{-1})/p}{\left[\text{Tr}(\hat{\Sigma}_n^{-1})/p\right]^2} - \frac{1}{\text{Tr}(\Sigma_n^{-1})/p}$$

Minimization problem:

$$\min_{\hat{\Delta}_n} \mathcal{L}_{n}^{\text{MV}}(U_n\hat{\Delta}_nU'_n, \Sigma_n)$$

Solution:

$$\Delta_n^* := \text{Diag}(\delta_{n,1}^*, \ldots, \delta_{n,p}^*) \quad \text{where} \quad \delta_{n,i}^* := u'_{n,i} \Sigma_n u_{n,i}$$

Note: Using the Frobenius loss instead yields the same solution.
Outline

1. The Problem
2. Finite Samples
3. Large-Dimensional Asymptotics
4. Kernel Estimation
5. Monte Carlo Study
6. Conclusion
Asymptotic Framework

Let \( p := p(n) \) and assume \( p/n \to c \in (0, 1) \), as \( n \to \infty \).

The following set of assumptions is maintained throughout:

A1 The population covariance matrix \( \Sigma_n \) is a nonrandom \( p \)-dimensional positive definite matrix.

A2 Let \( X_n \) be an \( n \times p \) matrix of real i.i.d. random variables with zero mean, unit variance, and finite 16th moment. One observes \( Y_n := X_n \Sigma_n^{1/2} \).

A3 Let \( \{(\tau_{n,1}, \ldots, \tau_{n,p}); (\nu_{n,1}, \ldots, \nu_{n,p})\} \) denote the eigenvalues and eigenvectors of \( \Sigma_n \). The e.d.f. of the population eigenvalues, denoted by \( H_n \), converges weakly to some limit e.d.f. \( H \).

A4 \( \text{Supp}(H) \), the support of \( H \), is the union of a finite number of closed intervals, bounded away from zero and infinity. Furthermore, there exists a compact interval in \((0, +\infty)\) which contains \( \text{Supp}(H_n) \) for all large enough \( n \).

Note: The paper also discusses an extension to the case \( p > 1 \).
Random Matrix Theory

A foundational result going back to Marčenko and Pastur (1967) states that the limiting distribution of the sample eigenvalues is deterministic.

Under the stated assumptions, there exists a continuous limiting sample spectral distribution $F$ such that $\forall x \in \mathbb{R} \ F_n(x) \xrightarrow{a.s.} F(x)$.

The limiting sample spectral c.d.f. $F$ is uniquely determined by $c$ and $H$; thus, we will refer to it as $F_{c,H} := F$ whenever clarification is needed.

A further implication is that the support of $F$ is the union of a finite number of compact intervals.
Illustration

$H$ is a point mass at one (such as for the identity covariance matrix).

Plot the density of $F$ for various values of $c$:
Hilbert Transform

The Hilbert transform of a real function $g$ is defined as

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

where $PV$ denotes the Cauchy Principal Value.

It is thus the convolution of $g$ with the Cauchy kernel $\frac{dt}{\pi t}$.

(Since the Cauchy kernel is singular, the integral does not converge in the usual sense and recourse to the Cauchy Principal Value is needed.)
Hilbert Transform

The Hilbert transform of a real function $g$ is defined as

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

where $\text{PV}$ denotes the Cauchy Principal Value.

It is thus the convolution of $g$ with the Cauchy kernel $\frac{dt}{\pi t}$.

(Since the Cauchy kernel is singular, the integral does not converge in the usual sense and recourse to the Cauchy Principal Value is needed.)

Intuition:

- The Hilbert transform operates like a local attraction force
- It pushes $x$ towards local mass centers
- For an illustration, plot the Hilbert transform of four densities
Illustration

- Uniform
- Cauchy
- Semicircle
- Arcsine

Density
Hilbert Transform
Optimal Nonlinear Shrinkage Formula

In our class of estimators, we can think of $\hat{\delta}_{n,i}$ as $\hat{\delta}_n(\lambda_{n,i})$, where $\hat{\delta}_n$ is an unrestricted nonlinear shrinkage function, assumed to converge to a limit $\delta$. 
Optimal Nonlinear Shrinkage Formula

In our class of estimators, we can think of \( \hat{\delta}_{n,i} \) as \( \hat{\delta}_n(\lambda_{n,i}) \), where \( \hat{\delta}_n \) is an unrestricted nonlinear shrinkage function, assumed to converge to a limit \( \hat{\delta} \).

Under the stated assumptions:

- For any \( \hat{\delta} \), the limiting loss of the estimator \( \hat{\Sigma}_n \) is deterministic.
- One can minimize this deterministic limiting loss wrt \( \hat{\delta} \).
- The solution yields an oracle nonlinear shrinkage formula.
Optimal Nonlinear Shrinkage Formula

In our class of estimators, we can think of \( \tilde{\delta}_{n,i} \) as \( \tilde{\delta}_n(\lambda_{n,i}) \), where \( \tilde{\delta}_n \) is an unrestricted nonlinear shrinkage function, assumed to converge to a limit \( \tilde{\delta} \).

Under the stated assumptions:

- For any \( \tilde{\delta} \), the limiting loss of the estimator \( \tilde{\Sigma}_n \) is deterministic
- One can minimize this deterministic limiting loss wrt \( \tilde{\delta} \)
- The solution yields an oracle nonlinear shrinkage formula

The oracle formula is given by

\[
\forall x \in \text{Supp}(F) \quad \delta^o(x) := \frac{x}{\left[ \pi cx f(x) \right]^2 + \left[ 1 - c - \pi cx H_f(x) \right]^2}
\]

where \( f \) denotes the density of \( F \).
Nonlinear Shrinkage as Local Attraction

The oracle formula results in **local attraction**: any sample eigenvalue is moved towards local mass centers.
Nonlinear Shrinkage as Local Attraction

The oracle formula results in local attraction: any sample eigenvalue is moved towards local mass centers.

This phenomenon is easier to see based on the ‘scaled’ density \( \varphi(x) := \pi x f(x) \), which yields the equivalent oracle formula

\[
\forall x \in \text{Supp}(F) \quad \delta^o(x) = \frac{x}{1 + c^2[\varphi^2(x) + H^2_{\varphi}(x)] - 2cH_{\varphi}(x)}
\]
Nonlinear Shrinkage as Local Attraction

The oracle formula results in **local attraction**: any sample eigenvalue is moved towards local mass centers.

This phenomenon is easier to see based on the ‘scaled’ density \( \varphi(x) := \pi x f(x) \), which yields the equivalent oracle formula

\[
\forall x \in \text{Supp}(F) \quad \delta^o(x) = \frac{x}{1 + c^2[\varphi^2(x) + H_x^2(\varphi)] - 2cH_x(\varphi)}
\]

**Crucial advantage** over global, linear shrinkage:

- Sample eigenvalues may be moved **away** from the grand mean, towards a local mass center
- It is helpful to consider an illustration:
  - \( H \) is a two-point mass at \( \{0.8, 2.0\} \), \( n = 18,000 \), and \( p = 4,000 \)
Illustration

Nonlinear Shrinkage with the Hilbert Transform

- **No Shrinkage**
- **Linear Shrinkage**
- **Nonlinear Shrinkage**

Shrunk Eigenvalues vs. Sample Eigenvalues

Histogram of Sample Eigenvalues
The oracle estimator $\hat{\Sigma}_n^o := U_n\Delta^o U'_n$ is not available in practice.

A bona fide estimator that also minimizes the asymptotic loss could be obtained via uniformly consistent estimation of $\delta^o$. 
Previous Approaches

QuEST:

- Indirect estimation of $\delta^o$
- Proposed by Ledoit and Wolf (2012, AoS; 2015, JMVA)
- First find consistent estimator $\hat{H}_n$ of $H$
- Then feed $\hat{H}_n$ into the Marčenko-Pastur equation, together with $\hat{c}_n := p/n$, and make use of the resulting $\hat{F}_n$
- Difficult to implement and slow to execute
- Cannot go much beyond dimension $p = 1000$ computationally
Previous Approaches

QuEST:
- Indirect estimation of $\delta^0$
- Proposed by Ledoit and Wolf (2012, AoS; 2015, JMVA)
- First find consistent estimator $\hat{H}_n$ of $H$
- Then feed $\hat{H}_n$ into the Marčenko-Pastur equation, together with $\hat{c}_n := p/n$, and make use of the resulting $\hat{F}_n$
- Difficult to implement and slow to execute
- Cannot go much beyond dimension $p = 1000$ computationally

NERCOME:
- Proposed by Abadir et al. (2104, JoE) and Lam (2106, AoS)
- Based on repeated sample splits to estimate the two components of $\delta^*_n := u'_{n,i} \Sigma_n u_{n,i}$ separately
- Requires brute-force spectral decomposition of many matrices
- Easy to implement but also slow to execute
- Cannot go much beyond dimension $p = 1000$ computationally
New Approach: Direct Estimation

Recall:

\[
\forall x \in \text{Supp}(F) \quad \delta^o(x) := \frac{x}{\left[\pi cx f(x)\right]^2 + \left[1 - c - \pi cx H_f(x)\right]^2}
\]

Therefore, **uniformly consistent estimation of \( \delta^o \) can be based on:**

(i) consistent estimation of \( c \)
(ii) uniformly consistent estimation of \( f \)
(iii) uniformly consistent estimation of \( H_f \)
New Approach: Direct Estimation

Recall:

\[
\forall x \in \text{Supp}(F) \quad \delta^0(x) := \frac{x}{\left[\pi cx f(x)\right]^2 + \left[1 - c - \pi cx H_f(x)\right]^2}
\]

Therefore, uniformly consistent estimation of \( \delta^0 \) can be based on:

(i) consistent estimation of \( c \)
(ii) uniformly consistent estimation of \( f \)
(iii) uniformly consistent estimation of \( H_f \)

Problem (i) is trivially solved by using \( \hat{c}_n := p/n \).

Problems (ii) and (iii) can be solved by kernel estimation.
Recall:

\[ \forall x \in \text{Supp}(F) \quad \delta^o(x) := \frac{x}{\left[ \pi c x f(x) \right]^2 + \left[ 1 - c - \pi c x H_f(x) \right]^2} \]

Therefore, uniformly consistent estimation of \( \delta^o \) can be based on:

(i) consistent estimation of \( c \)
(ii) uniformly consistent estimation of \( f \)
(iii) uniformly consistent estimation of \( H_f \)

Problem (i) is trivially solved by using \( \hat{c}_n := p/n \).

Problems (ii) and (iii) can be solved by kernel estimation.

Advantages:
- Easy to implement and fast to execute
- Can go to at least dimension \( p = 10,000 \) computationally
Outline

1. The Problem
2. Finite Samples
3. Large-Dimensional Asymptotics
4. Kernel Estimation
5. Monte Carlo Study
6. Conclusion
A kernel $k(\cdot)$ is assumed to satisfy the following properties:

- $k$ is a continuous, symmetric density with finite support, mean zero, and variance one.
- Its Hilbert transform $\mathcal{H}_k$ exists and is continuous.
- Both the kernel $k$ and its Hilbert transform $\mathcal{H}_k$ are functions of bounded variation.
Choice of Kernel

A kernel $k(\cdot)$ is assumed to satisfy the following properties:

- $k$ is a continuous, symmetric density with finite support, mean zero, and variance one
- Its Hilbert transform $\mathcal{H}_k$ exists and is continuous
- Both the kernel $k$ and its Hilbert transform $\mathcal{H}_k$ are functions of bounded variation

We use the semi-circle kernel dating back to Wigner (1955, AoM).
Choice of Kernel

A kernel $k(\cdot)$ is assumed to satisfy the following properties:

- $k$ is a continuous, symmetric density with finite support, mean zero, and variance one
- Its Hilbert transform $H_k$ exists and is continuous
- Both the kernel $k$ and its Hilbert transform $H_k$ are functions of bounded variation

We use the semi-circle kernel dating back to Wigner (1955, AoM).

Of the 48 elementary functions whose Hilbert transform is known in closed form, it is the only one satisfying all the above assumptions.
We propose to use a variable bandwidth that is proportional to the magnitude of a given sample eigenvalue.

The bandwidth applied to $\lambda_{n,i}$ is $h_{n,i} := \lambda_{n,i} h_n$, where $h_n \to 0$.

We use $h_n := n^{-0.35}$, close to the choice $n^{-1/3}$ by Jing et al. (2010, AoS). (Although they use a uniform bandwidth $h_{n,i} \equiv n^{-1/3}$).
Kernel Estimators & Feasible Shrinkage Formula

Kernel estimators of $f$ and $H_f$:

$$\forall x \in \mathbb{R} \quad \tilde{f}_n(x) := \frac{1}{p} \sum_{i=1}^{p} \frac{1}{h_{n,i}} k \left( \frac{x - \lambda_{n,i}}{h_{n,i}} \right)$$
Kernel Estimators & Feasible Shrinkage Formula

Kernel estimators of $f$ and $H_f$:

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

$\forall x \in \mathbb{R} \quad H_{\tilde{f}_n}(x) := \frac{1}{p} \sum_{i=1}^{p} \frac{1}{h_{n,i}} H_k\left(\frac{x - \lambda_{n,i}}{h_{n,i}}\right) = \frac{1}{\pi} PV \int \frac{\tilde{f}_n(t)}{x-t} dt$
Kernel Estimators & Feasible Shrinkage Formula

Kernel estimators of $f$ and $H_f$:

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

\[
\forall x \in \mathbb{R} \quad \tilde{H}_{f_n}(x) := \frac{1}{p} \sum_{i=1}^{p} \frac{1}{h_{n,i}} \mathcal{H}_k \left( \frac{x - \lambda_{n,i}}{h_{n,i}} \right) = \frac{1}{\pi} \text{PV} \int \frac{\tilde{f}_n(t)}{x-t} dt
\]

Feasible nonlinear shrinkage estimation:

\[
\forall x \in \text{Supp}(F) \quad \delta_n(x) := \frac{x}{\left[ \pi c_n x \tilde{f}_n(x) \right]^2 + \left[ 1 - c_n - \pi c_n x \tilde{H}_{f_n}(x) \right]^2}
\]
Kernel Estimators & Feasible Shrinkage Formula

Kernel estimators of $f$ and $H_f$:

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

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

Feasible nonlinear shrinkage estimation:

$$\forall x \in \text{Supp}(F) \quad \tilde{\delta}_n(x) := \frac{x}{\left[ \pi \tilde{c}_n x \tilde{f}_n(x) \right]^2 + \left[ 1 - \tilde{c}_n - \pi \tilde{c}_n x \mathcal{H}_{\tilde{f}_n}(x) \right]^2}$$

$$\tilde{\Sigma}_n := U_n \tilde{\Delta}_n U_n'$$
Outline

1. The Problem
2. Finite Samples
3. Large-Dimensional Asymptotics
4. Kernel Estimation
5. Monte Carlo Study
6. Conclusion
Executive Summary

Performance of direct nonlinear shrinkage:

- Much better than linear shrinkage
- Basically as good as QuEST
- Somewhat better than NERCOME
Executive Summary

**Performance** of direct nonlinear shrinkage:
- Much better than linear shrinkage
- Basically as good as QuEST
- Somewhat better than NERCOME

**Speed** of direct nonlinear shrinkage:
- Basically as fast as linear shrinkage
- Much faster than QuEST
- Much faster than NERCOME
Executive Summary

**Performance** of direct nonlinear shrinkage:
- Much better than linear shrinkage
- Basically as good as QuEST
- Somewhat better than NERCOME

**Speed** of direct nonlinear shrinkage:
- Basically as fast as linear shrinkage
- Much faster than QuEST
- Much faster than NERCOME

⇒ Get the best of both worlds!
Main Performance Measure

**Percentage Relative Improvement in Average Loss (PRIAL):**

\[
PRIAL_{MV}^n(\hat{\Sigma}_n) := \frac{\mathbb{E}[L_{MV}^n(S_n, \Sigma_n)] - \mathbb{E}[L_{MV}^n(\hat{\Sigma}_n, \Sigma_n)]}{\mathbb{E}[L_{MV}^n(S_n, \Sigma_n)] - \mathbb{E}[L_{MV}^n(S^*_n, \Sigma_n)]} \times 100\%
\]
Main Performance Measure

Percentage Relative Improvement in Average Loss (PRIAL):

\[
PRIAL_{n}^{MV}(\Sigma_n) := \frac{\mathbb{E}[\mathcal{L}_{n}^{MV}(S_n, \Sigma_n)] - \mathbb{E}[\mathcal{L}_{n}^{MV}(\hat{\Sigma}_n, \Sigma_n)]}{\mathbb{E}[\mathcal{L}_{n}^{MV}(S_n, \Sigma_n)] - \mathbb{E}[\mathcal{L}_{n}^{MV}(S^*_n, \Sigma_n)]} \times 100\%
\]

By construction:

- Sample covariance matrix has \(PRIAL_{n}^{MV}(S_n) = 0\%
- Finite-sample optimal estimator has \(PRIAL_{n}^{MV}(S^*_n) = 100\%\)
Main Performance Measure

Percentage Relative Improvement in Average Loss (PRIAL):

\[
PRIAL_{n}^{MV}(\hat{\Sigma}) := \frac{\mathbb{E}[L_{n}^{MV}(S, \Sigma)] - \mathbb{E}[L_{n}^{MV}(\hat{\Sigma}, \Sigma)]}{\mathbb{E}[L_{n}^{MV}(S, \Sigma)] - \mathbb{E}[L_{n}^{MV}(\hat{S}, \Sigma)]} \times 100\%
\]

By construction:

- Sample covariance matrix has \(PRIAL_{n}^{MV}(S) = 0\%\)
- Finite-sample optimal estimator has \(PRIAL_{n}^{MV}(\hat{S}) = 100\%\)

Note:

- Negative PRIAL values are possible
We use a scenario introduced by Bai and Silverstein (1998, AoP):

- Dimension $p = 200$
- Sample size $n = 600$
- Concentration ratio $\hat{c}_n = 1/3$
- 20% of the $\tau_{n,i}$ are equal to 1, 40% equal to 3, and 40% equal to 10
- Condition number $\theta = 10$
- Variates are normally distributed
Baseline Scenario

We use a scenario introduced by Bai and Silverstein (1998, AoP):

- Dimension $p = 200$
- Sample size $n = 600$
- Concentration ratio $\hat{c}_n = 1/3$
- 20% of the $\tau_{n,i}$ are equal to 1, 40% equal to 3, and 40% equal to 10
- Condition number $\theta = 10$
- Variates are normally distributed

Each feature will be varied in subsequent scenarios.
## Results for Baseline Scenario

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Sample</th>
<th>Linear</th>
<th>Direct</th>
<th>QuEST</th>
<th>NERCOME</th>
<th>FSOPT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ø Loss</td>
<td>2.71</td>
<td>2.10</td>
<td>1.52</td>
<td>1.50</td>
<td>1.58</td>
<td>1.48</td>
</tr>
<tr>
<td>PRIAL</td>
<td>0%</td>
<td>50%</td>
<td>97%</td>
<td>98%</td>
<td>92%</td>
<td>100%</td>
</tr>
<tr>
<td>Time (ms)</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>2,233</td>
<td>2,990</td>
<td>3</td>
</tr>
</tbody>
</table>

Note:
- Computational times in milliseconds come from a 64-bit, quad-core 4.00GHz Windows PC running Matlab R2016a
Large-Dimensional Asymptotics

Let $p$ and $n$ go to infinity together with $p/n \equiv 1/3$: 
Large-Dimensional Asymptotics

Let $p$ and $n$ go to infinity together with $p/n \equiv 1/3$:

Monte Carlo Simulations: Convergence

- QuEST
- Direct
- NERCOME
- Linear
Speed

Let $p$ and $n$ go to infinity together with $p/n \equiv 1/3$: 
Let $p$ and $n$ go to infinity together with $p/n \equiv 1/3$:
Ultra-High Dimension

Repeat baseline scenario but multiply both $p$ and $n$ by 50:

- $p = 10,000$
- $n = 30,000$

QuEST and NERCOME are no longer computationally feasible.
Repeat baseline scenario but multiply both $p$ and $n$ by 50:

- $p = 10,000$
- $n = 30,000$

QuEST and NERCOME are no longer computationally feasible.

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Sample</th>
<th>Linear</th>
<th>Direct</th>
<th>FSOPT</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\emptyset$ Loss</td>
<td>2.679</td>
<td>2.086</td>
<td>1.488</td>
<td>1.487</td>
</tr>
<tr>
<td>PRIAL</td>
<td>0%</td>
<td>49.74%</td>
<td>99.92%</td>
<td>100%</td>
</tr>
<tr>
<td>Time (s)</td>
<td>21</td>
<td>43</td>
<td>113</td>
<td>108</td>
</tr>
</tbody>
</table>
Concentration Ratio

Vary $p/n$ from 0.1 to 0.9 while keeping $p \times n = 120,000$: 
Concentration Ratio

Vary $p/n$ from 0.1 to 0.9 while keeping $p \times n = 120,000$:

Monte Carlo Simulations: Concentration

PRIAL (%)
Condition Number

Vary $\theta$ from 3 to 30, by linearly squeezing/stretching the $\tau_{n,i}$:
Vary $\theta$ from 3 to 30, by linearly squeezing/stretching the $\tau_{n,i}$:

Monte Carlo Simulations: Condition

- QuEST
- Direct
- NERCOME
- Linear
Non-Normality

Vary the distribution of the variates:
## Non-Normality

Vary the distribution of the variates:

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Linear</th>
<th>Direct</th>
<th>QuEST</th>
<th>NERCOME</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>50%</td>
<td>97%</td>
<td>98%</td>
<td>92%</td>
</tr>
<tr>
<td>Bernoulli</td>
<td>51%</td>
<td>98%</td>
<td>98%</td>
<td>92%</td>
</tr>
<tr>
<td>Laplace</td>
<td>50%</td>
<td>97%</td>
<td>98%</td>
<td>92%</td>
</tr>
<tr>
<td>‘Student’ $t_5$</td>
<td>49%</td>
<td>97%</td>
<td>98%</td>
<td>92%</td>
</tr>
</tbody>
</table>
Use a shifted and stretched Beta distribution with support $[1,10]$:
Use a shifted and stretched Beta distribution with support $[1,10]$:

<table>
<thead>
<tr>
<th>Beta Parameters</th>
<th>Linear</th>
<th>Direct</th>
<th>QuEST</th>
<th>NERCOME</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1, 1)</td>
<td>83%</td>
<td>98%</td>
<td>99%</td>
<td>96%</td>
</tr>
<tr>
<td>(1, 2)</td>
<td>95%</td>
<td>99%</td>
<td>99%</td>
<td>98%</td>
</tr>
<tr>
<td>(2, 1)</td>
<td>94%</td>
<td>99%</td>
<td>99%</td>
<td>99%</td>
</tr>
<tr>
<td>(1.5, 1.5)</td>
<td>92%</td>
<td>99%</td>
<td>99%</td>
<td>98%</td>
</tr>
<tr>
<td>(0.5, 0.5)</td>
<td>50%</td>
<td>98%</td>
<td>98%</td>
<td>94%</td>
</tr>
<tr>
<td>(5, 5)</td>
<td>98%</td>
<td>100%</td>
<td>100%</td>
<td>99%</td>
</tr>
<tr>
<td>(5, 2)</td>
<td>97%</td>
<td>100%</td>
<td>100%</td>
<td>98%</td>
</tr>
<tr>
<td>(2, 5)</td>
<td>99%</td>
<td>99%</td>
<td>99%</td>
<td>99%</td>
</tr>
</tbody>
</table>
Shape

Selected (shifted and stretched) beta densities used:

Shape of the Beta Density for Various Parameters ($\alpha, \beta$)

- $\times$ (1,1)
- $\bullet$ (1,2)
- $\ldots$ (1.5,1.5)
- $\ldots$ (0.5,0.5)
- $\bigcirc$ (5,5)
- $\bigcirc$ (2,5)
Fixed-Dimensional Asymptotics

Let $n$ grow from 250 to 20,000 while keeping $p \equiv 200$: 
Fixed-Dimensional Asymptotics

Let $n$ grow from 250 to 20,000 while keeping $p = 200$:

Monte Carlo Simulations: FixedDim

![Graph showing Monte Carlo simulations for different methods. The x-axis represents sample size on a log scale, and the y-axis represents PRIAL (%). Different line styles and colors represent QuEST, Direct, NERCOME, and Linear methods.]
Arrow Model

Let $\tau_{n,p} := 1 + 0.5(p - 1)$ and remaining bulk from s&s Beta(5,2):
Let $\tau_{n,p} := 1 + 0.5(p - 1)$ and remaining bulk from s&s Beta(5,2):
Nonlinear shrinkage estimation of covariance matrices is a complex, but powerful structure-free approach in large dimensions.

Existing methods are difficult to implement, computationally expensive, or even both.

We have suggested a direct method based on kernel estimation that (i) performs as well as existing methods and (ii) is computationally as cheap as linear shrinkage.
Nonlinear shrinkage estimation of covariance matrices is a complex, but powerful structure-free approach in large dimensions.

Existing methods are difficult to implement, computationally expensive, or even both.

We have suggested a direct method based on kernel estimation that (i) performs as well as existing methods and (ii) is computationally as cheap as linear shrinkage.

This direct method also can handle dimensions of +1 magnitude, which is a big + in the age of Big Data.


