Analytical Nonlinear Shrinkage of Large-Dimensional Covariance Matrices

Olivier Ledoit[1] and Michael Wolf[1]

[1]Department of Economics
University of Zurich

RMCDA Shanghai, December 11th, 2019

Outline

Outline

What is the Point of the Paper?

What is the Point of the Paper?

To solve with random matrix theory a very general statistical problem

What is the Point of the Paper?

To solve with random matrix theory a very general statistical problem

How to Estimate the Covariance Matrix

"the *second* most important object in all of Statistics"

What is the Point of the Paper?

To solve with random matrix theory a very general statistical problem

How to Estimate the Covariance Matrix

"the *second* most important object in all of Statistics"

How do we do it?

What is the Point of the Paper?

To solve with random matrix theory a very general statistical problem

How to Estimate the Covariance Matrix

"the *second* most important object in all of Statistics"

How do we do it?

By combining Olivier Ledoit and Sandrine Péché (2011) with Bing-Yi Jing, Guangming Pan, Qi-Man Shao and Wang Zhou (2010).

Many Applications besides Finance

Many Applications besides Finance

- cancer research (Pyeon et al., 2007)
- chemistry (Guo et al., 2012)
- civil engineering (Michaelides et al., 2011)
- climatology (Ribes et al., 2009)
- electrical engineering (Wei et al., 2011)
- genetics (Lin et al., 2012)
- geology (Elsheikh et al., 2013)
- neuroscience (Fritsch et al., 2012)
- psychology (Markon, 2010)
- speech recognition (Bell and King, 2009)
- etc...

Overall Plan of the Talk

Overall Plan of the Talk

1. Set up required background in Multivariate Statistics

Overall Plan of the Talk

1. Set up required background in Multivariate Statistics
2. Review useful results from Random Matrix Theory

Overall Plan of the Talk

1. Set up required background in Multivariate Statistics
2. Review useful results from Random Matrix Theory
3. Bring both threads together by estimating a Hilbert transform

Overall Plan of the Talk

1. Set up required background in Multivariate Statistics
2. Review useful results from Random Matrix Theory
3. Bring both threads together by estimating a Hilbert transform
4. Report Monte Carlo simulations

Overall Plan of the Talk

1. Set up required background in Multivariate Statistics
2. Review useful results from Random Matrix Theory
3. Bring both threads together by estimating a Hilbert transform
4. Report Monte Carlo simulations
5. Run empirical experiment on real-world financial data

Outline

The Sample Covariance Matrix

- Y_n: matrix of n iid observations on p zero-mean variables
- Sample covariance matrix $S_n := Y_n' Y_n / n$
- Population covariance matrix $\Sigma_n := \mathbb{E}[S_n]$

The Sample Covariance Matrix

- Y_n: matrix of n iid observations on p zero-mean variables
- Sample covariance matrix $S_n := Y_n' Y_n / n$
- Population covariance matrix $\Sigma_n := \mathbb{E}[S_n]$

- Problem 1: S_n is non-invertible when $p > n$
- Problem 2: S_n is ill-conditioned when n is not much bigger than p
- Problem 3: S_n is *inadmissible* when $p \geq 3$ (James and Stein, 1961)

[Inadmissible means that there exists a more accurate estimator.]

Class of Rotation-Equivariant Estimators

Class of Rotation-Equivariant Estimators

A Reasonable Request

Class of Rotation-Equivariant Estimators

A Reasonable Request

- $\widehat{\Sigma}_n := \widehat{\Sigma}_n(Y_n)$ is a generic estimator of Σ_n

Class of Rotation-Equivariant Estimators

A Reasonable Request

- $\widehat{\Sigma}_n := \widehat{\Sigma}_n(Y_n)$ is a generic estimator of Σ_n
- R is a $p \times p$ rotation matrix: $R^{-1} = R'$

Class of Rotation-Equivariant Estimators

A Reasonable Request

- $\widehat{\Sigma}_n := \widehat{\Sigma}_n(Y_n)$ is a generic estimator of Σ_n
- R is a $p \times p$ rotation matrix: $R^{-1} = R'$
- Rotation equivariance means $\widehat{\Sigma}_n(Y_n R) = R' \widehat{\Sigma}_n(Y_n) R$

Class of Rotation-Equivariant Estimators

A Reasonable Request

- $\widehat{\Sigma}_n := \widehat{\Sigma}_n(Y_n)$ is a generic estimator of Σ_n
- R is a $p \times p$ rotation matrix: $R^{-1} = R'$
- Rotation equivariance means $\widehat{\Sigma}_n(Y_n R) = R' \widehat{\Sigma}_n(Y_n) R$

No *a priori* information on orientation of population eigenvectors

Class of Rotation-Equivariant Estimators

A Reasonable Request

- $\widehat{\Sigma}_n := \widehat{\Sigma}_n(Y_n)$ is a generic estimator of Σ_n
- R is a $p \times p$ rotation matrix: $R^{-1} = R'$
- Rotation equivariance means $\widehat{\Sigma}_n(Y_n R) = R' \widehat{\Sigma}_n(Y_n) R$

No *a priori* information on orientation of population eigenvectors

Stein (1986) shows it is the same as keeping the sample eigenvectors and modifying the sample eigenvalues:

Class of Rotation-Equivariant Estimators

A Reasonable Request

- $\widehat{\Sigma}_n := \widehat{\Sigma}_n(Y_n)$ is a generic estimator of Σ_n
- R is a $p \times p$ rotation matrix: $R^{-1} = R'$
- Rotation equivariance means $\widehat{\Sigma}_n(Y_n R) = R' \widehat{\Sigma}_n(Y_n) R$

No *a priori* information on orientation of population eigenvectors

Stein (1986) shows it is the same as keeping the sample eigenvectors and modifying the sample eigenvalues:

$\lambda_{n,1}, \ldots, \lambda_{n,p}$: sample eigenvalues; $u_{n,1}, \ldots, u_{n,p}$: sample eigenvectors

Class of Rotation-Equivariant Estimators

A Reasonable Request

- $\widehat{\Sigma}_n := \widehat{\Sigma}_n(Y_n)$ is a generic estimator of Σ_n
- R is a $p \times p$ rotation matrix: $R^{-1} = R'$
- Rotation equivariance means $\widehat{\Sigma}_n(Y_n R) = R' \widehat{\Sigma}_n(Y_n) R$

No *a priori* information on orientation of population eigenvectors

Stein (1986) shows it is the same as keeping the sample eigenvectors and modifying the sample eigenvalues:

$\lambda_{n,1}, \dots, \lambda_{n,p}$: sample eigenvalues; $u_{n,1}, \dots, u_{n,p}$: sample eigenvectors

$$S_n = \sum_{i=1}^{p} \lambda_{n,i} \cdot u_{n,i} u'_{n,i} \quad \longrightarrow \quad \widehat{\Sigma}_n = \sum_{i=1}^{p} \widehat{\delta}_{n,i} \cdot u_{n,i} u'_{n,i}$$

Comparison with Other Approaches

Comparison with Other Approaches

(1) This is not the *sparsity* approach of Bickel and Levina (2008, AoS):

Comparison with Other Approaches

(1) This is not the *sparsity* approach of Bickel and Levina (2008, AoS):

- they assume that the current orthormal basis is special in the sense that most of the $p(p-1)/2$ covariances are zero

Comparison with Other Approaches

(1) This is not the *sparsity* approach of Bickel and Levina (2008, AoS):

- they assume that the current orthormal basis is special in the sense that most of the $p(p-1)/2$ covariances are zero
- this condition would not hold true in general after rotation

Comparison with Other Approaches

(1) This is not the *sparsity* approach of Bickel and Levina (2008, AoS):

- they assume that the current orthormal basis is special in the sense that most of the $p(p-1)/2$ covariances are zero
- this condition would not hold true in general after rotation
- it requires *a priori* information about the orientation of the eigenvectors of the population covariance matrix, which is unverifiable in practice.

Comparison with Other Approaches

(1) This is not the *sparsity* approach of Bickel and Levina (2008, AoS):

- they assume that the current orthormal basis is special in the sense that most of the $p(p-1)/2$ covariances are zero
- this condition would not hold true in general after rotation
- it requires *a priori* information about the orientation of the eigenvectors of the population covariance matrix, which is unverifiable in practice.

(2) This is not the *linear shrinkage* of Ledoit and Wolf (2004, JMVA):

Comparison with Other Approaches

(1) This is not the *sparsity* approach of Bickel and Levina (2008, AoS):

- they assume that the current orthormal basis is special in the sense that most of the $p(p-1)/2$ covariances are zero
- this condition would not hold true in general after rotation
- it requires *a priori* information about the orientation of the eigenvectors of the population covariance matrix, which is unverifiable in practice.

(2) This is not the *linear shrinkage* of Ledoit and Wolf (2004, JMVA):

- they assume the modified eigenvalues are linear functions of the observed ones: $\forall i = 1, \ldots, p \quad \widehat{\delta}_{n,i} = a_n + b_n \lambda_{n,i}$

Comparison with Other Approaches

(1) This is not the *sparsity* approach of Bickel and Levina (2008, AoS):

- they assume that the current orthormal basis is special in the sense that most of the $p(p-1)/2$ covariances are zero
- this condition would not hold true in general after rotation
- it requires *a priori* information about the orientation of the eigenvectors of the population covariance matrix, which is unverifiable in practice.

(2) This is not the *linear shrinkage* of Ledoit and Wolf (2004, JMVA):

- they assume the modified eigenvalues are linear functions of the observed ones: $\forall i = 1, \ldots, p \quad \widehat{\delta}_{n,i} = a_n + b_n \lambda_{n,i}$
- they have only 2 degrees of freedom, whereas our class has $p \gg 2$ degrees of freedom

Comparison with Other Approaches

(1) This is not the *sparsity* approach of Bickel and Levina (2008, AoS):

- they assume that the current orthormal basis is special in the sense that most of the $p(p-1)/2$ covariances are zero
- this condition would not hold true in general after rotation
- it requires *a priori* information about the orientation of the eigenvectors of the population covariance matrix, which is unverifiable in practice.

(2) This is not the *linear shrinkage* of Ledoit and Wolf (2004, JMVA):

- they assume the modified eigenvalues are linear functions of the observed ones: $\forall i = 1, \ldots, p \quad \widehat{\delta}_{n,i} = a_n + b_n \lambda_{n,i}$
- they have only 2 degrees of freedom, whereas our class has $p \gg 2$ degrees of freedom
- linear shrinkage is a good first-order approximation if optimal nonlinear shrinkage happens to be 'almost' linear, but in the general case it can be further improved

Loss Functions

Loss Functions

Frobenius: $\mathcal{L}_n^{FR}\left(\widehat{\Sigma}_n, \Sigma_n\right) := \frac{1}{p}\mathrm{Tr}\left[\left(\widehat{\Sigma}_n - \Sigma_n\right)^2\right]$

Minimum Variance: $\mathcal{L}_{\mathbf{n}}^{\mathbf{MV}}\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{n}}, \boldsymbol{\Sigma}_{\mathbf{n}}\right) := \frac{\mathrm{Tr}\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{n}}^{-1}\boldsymbol{\Sigma}_{\mathbf{n}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{n}}^{-1}\right)/\mathbf{p}}{\left[\mathrm{Tr}\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{n}}^{-1}\right)/\mathbf{p}\right]^2} - \frac{\mathbf{1}}{\mathrm{Tr}\left(\boldsymbol{\Sigma}_{\mathbf{n}}^{-1}\right)/\mathbf{p}}$

Inverse Stein: $\mathcal{L}_n^{IS}\left(\widehat{\Sigma}_n, \Sigma_n\right) := \frac{1}{p}\mathrm{Tr}\left[\Sigma_n\widehat{\Sigma}_n^{-1}\right] - \frac{1}{p}\log\left[\det\left(\Sigma_n\widehat{\Sigma}_n^{-1}\right)\right]$

Stein: $\mathcal{L}_n^{ST}\left(\widehat{\Sigma}_n, \Sigma_n\right) := \frac{1}{p}\mathrm{Tr}\left[\Sigma_n^{-1}\widehat{\Sigma}_n\right] - \frac{1}{p}\log\left[\det\left(\Sigma_n^{-1}\widehat{\Sigma}_n\right)\right]$

Inverse Frobenius: $\mathcal{L}_n^{IF}\left(\widehat{\Sigma}_n, \Sigma_n\right) := \frac{1}{p}\mathrm{Tr}\left[\left(\widehat{\Sigma}_n^{-1} - \Sigma_n^{-1}\right)^2\right]$

Weighted Frobenius: $\mathcal{L}_n^{WF}\left(\widehat{\Sigma}_n, \Sigma_n\right) := \frac{1}{p}\mathrm{Tr}\left[\left(\widehat{\Sigma}_n - \Sigma_n\right)^2\Sigma_n^{-1}\right]$

Loss Functions

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

Minimum Variance: $\mathcal{L}_{\mathbf{n}}^{\mathbf{MV}}\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{n}}, \boldsymbol{\Sigma}_{\mathbf{n}}\right) := \dfrac{\text{Tr}\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{n}}^{-1}\boldsymbol{\Sigma}_{\mathbf{n}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{n}}^{-1}\right)/\mathbf{p}}{\left[\text{Tr}\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{n}}^{-1}\right)/\mathbf{p}\right]^2} - \dfrac{\mathbf{1}}{\text{Tr}\left(\boldsymbol{\Sigma}_{\mathbf{n}}^{-1}\right)/\mathbf{p}}$

Inverse Stein: $\mathcal{L}_n^{IS}\left(\widehat{\Sigma}_n, \Sigma_n\right) := \frac{1}{p}\text{Tr}\left[\Sigma_n\widehat{\Sigma}_n^{-1}\right] - \frac{1}{p}\log\left[\det\left(\Sigma_n\widehat{\Sigma}_n^{-1}\right)\right]$

Stein: $\mathcal{L}_n^{ST}\left(\widehat{\Sigma}_n, \Sigma_n\right) := \frac{1}{p}\text{Tr}\left[\Sigma_n^{-1}\widehat{\Sigma}_n\right] - \frac{1}{p}\log\left[\det\left(\Sigma_n^{-1}\widehat{\Sigma}_n\right)\right]$

Inverse Frobenius: $\mathcal{L}_n^{IF}\left(\widehat{\Sigma}_n, \Sigma_n\right) := \frac{1}{p}\text{Tr}\left[\left(\widehat{\Sigma}_n^{-1} - \Sigma_n^{-1}\right)^2\right]$

Weighted Frobenius: $\mathcal{L}_n^{WF}\left(\widehat{\Sigma}_n, \Sigma_n\right) := \frac{1}{p}\text{Tr}\left[\left(\widehat{\Sigma}_n - \Sigma_n\right)^2\Sigma_n^{-1}\right]$

We use the Minimum-Variance Loss championed by
Rob Engle, Olivier Ledoit and Michael Wolf (2019)

Finite-Sample Optimal (FSOPT) Estimator

Finite-Sample Optimal (FSOPT) Estimator

Find rotation-equivariant estimator closest to Σ_n according to MV loss

Finite-Sample Optimal (FSOPT) Estimator

Find rotation-equivariant estimator closest to Σ_n according to MV loss

Optimization problem:

$$\min_{\widehat{\delta}_{n,1},\dots,\widehat{\delta}_{n,p}} \mathcal{L}_n^{\mathrm{MV}}\left(\sum_{i=1}^{p} \widehat{\delta}_{n,i} \cdot u_{n,i}u_{n,i}', \Sigma_n\right)$$

Finite-Sample Optimal (FSOPT) Estimator

Find rotation-equivariant estimator closest to Σ_n according to MV loss

Optimization problem:

$$\min_{\widehat{\delta}_{n,1},\ldots,\widehat{\delta}_{n,p}} \mathcal{L}_n^{\mathrm{MV}}\left(\sum_{i=1}^{p} \widehat{\delta}_{n,i} \cdot u_{n,i}u_{n,i}', \Sigma_n\right)$$

Solution:

$$S_n^* := \sum_{i=1}^{p} d_{n,i}^* \cdot u_{n,i}u_{n,i}' \quad \text{where} \quad d_{n,i}^* := u_{n,i}'\Sigma_n u_{n,i} \tag{1}$$

Finite-Sample Optimal (FSOPT) Estimator

Find rotation-equivariant estimator closest to Σ_n according to MV loss

Optimization problem:

$$\min_{\widehat{\delta}_{n,1},\ldots,\widehat{\delta}_{n,p}} \mathcal{L}_n^{\mathrm{MV}}\left(\sum_{i=1}^{p} \widehat{\delta}_{n,i} \cdot u_{n,i}u'_{n,i}, \Sigma_n\right)$$

Solution:

$$S_n^* := \sum_{i=1}^{p} d_{n,i}^* \cdot u_{n,i}u'_{n,i} \quad \text{where} \quad d_{n,i}^* := u'_{n,i}\Sigma_n u_{n,i} \tag{1}$$

- Very intuitive: $d_{n,i}^*$ is the true variance of the linear combination of original variables weighted by eigenvector $u_{n,i}$

Finite-Sample Optimal (FSOPT) Estimator

Find rotation-equivariant estimator closest to Σ_n according to MV loss

Optimization problem:

$$\min_{\widehat{\delta}_{n,1},\ldots,\widehat{\delta}_{n,p}} \mathcal{L}_n^{\text{MV}}\left(\sum_{i=1}^{p} \widehat{\delta}_{n,i} \cdot u_{n,i}u_{n,i}', \Sigma_n\right)$$

Solution:

$$S_n^* := \sum_{i=1}^{p} d_{n,i}^* \cdot u_{n,i}u_{n,i}' \quad \text{where} \quad d_{n,i}^* := u_{n,i}'\Sigma_n u_{n,i} \tag{1}$$

- Very intuitive: $d_{n,i}^*$ is the true variance of the linear combination of original variables weighted by eigenvector $u_{n,i}$
- By contrast, $\lambda_{n,i}$ is the sample variance of the linear combination of original variables weighted by eigenvector $u_{n,i}$: overfitting!

Finite-Sample Optimal (FSOPT) Estimator

Find rotation-equivariant estimator closest to Σ_n according to MV loss

Optimization problem:

$$\min_{\widehat{\delta}_{n,1},\ldots,\widehat{\delta}_{n,p}} \mathcal{L}_n^{\mathrm{MV}}\left(\sum_{i=1}^{p} \widehat{\delta}_{n,i} \cdot u_{n,i}u'_{n,i}, \Sigma_n\right)$$

Solution:

$$S_n^* := \sum_{i=1}^{p} d_{n,i}^* \cdot u_{n,i}u'_{n,i} \quad \text{where} \quad d_{n,i}^* := u'_{n,i}\Sigma_n u_{n,i} \tag{1}$$

- Very intuitive: $d_{n,i}^*$ is the true variance of the linear combination of original variables weighted by eigenvector $u_{n,i}$
- By contrast, $\lambda_{n,i}$ is the sample variance of the linear combination of original variables weighted by eigenvector $u_{n,i}$: overfitting!
- FSOPT is the unattainable **'Gold Standard'**

A Numerical Scheme: NERCOME

A Numerical Scheme: NERCOME

- Proposed by Abadir et al. (2014) and Lam (2016, AoS)

A Numerical Scheme: NERCOME

- Proposed by Abadir et al. (2014) and Lam (2016, AoS)
- Split the sample into two parts

A Numerical Scheme: NERCOME

- Proposed by Abadir et al. (2014) and Lam (2016, AoS)
- Split the sample into two parts
- Estimate the eigenvectors from the first part

A Numerical Scheme: NERCOME

- Proposed by Abadir et al. (2014) and Lam (2016, AoS)
- Split the sample into two parts
- Estimate the eigenvectors from the first part
- Estimate the $d^*_{n,i}$'s from the second part

A Numerical Scheme: NERCOME

- Proposed by Abadir et al. (2014) and Lam (2016, AoS)
- Split the sample into two parts
- Estimate the eigenvectors from the first part
- Estimate the $d^*_{n,i}$'s from the second part
- Average over many different ways to split the sample

A Numerical Scheme: NERCOME

- Proposed by Abadir et al. (2014) and Lam (2016, AoS)
- Split the sample into two parts
- Estimate the eigenvectors from the first part
- Estimate the $d^*_{n,i}$'s from the second part
- Average over many different ways to split the sample
- Gets around the overfitting problem

A Numerical Scheme: NERCOME

- Proposed by Abadir et al. (2014) and Lam (2016, AoS)
- Split the sample into two parts
- Estimate the eigenvectors from the first part
- Estimate the $d^*_{n,i}$'s from the second part
- Average over many different ways to split the sample
- Gets around the overfitting problem

Problems:

A Numerical Scheme: NERCOME

- Proposed by Abadir et al. (2014) and Lam (2016, AoS)
- Split the sample into two parts
- Estimate the eigenvectors from the first part
- Estimate the $d^*_{n,i}$'s from the second part
- Average over many different ways to split the sample
- Gets around the overfitting problem

Problems:

- Requires brute-force spectral decomposition of many matrices

A Numerical Scheme: NERCOME

- Proposed by Abadir et al. (2014) and Lam (2016, AoS)
- Split the sample into two parts
- Estimate the eigenvectors from the first part
- Estimate the $d^*_{n,i}$'s from the second part
- Average over many different ways to split the sample
- Gets around the overfitting problem

Problems:

- Requires brute-force spectral decomposition of many matrices
- Easy to code but slow to execute

A Numerical Scheme: NERCOME

- Proposed by Abadir et al. (2014) and Lam (2016, AoS)
- Split the sample into two parts
- Estimate the eigenvectors from the first part
- Estimate the $d^*_{n,i}$'s from the second part
- Average over many different ways to split the sample
- Gets around the overfitting problem

Problems:

- Requires brute-force spectral decomposition of many matrices
- Easy to code but slow to execute
- Cannot go much beyond dimension $p = 1000$ computationally

A Numerical Scheme: NERCOME

- Proposed by Abadir et al. (2014) and Lam (2016, AoS)
- Split the sample into two parts
- Estimate the eigenvectors from the first part
- Estimate the $d^*_{n,i}$'s from the second part
- Average over many different ways to split the sample
- Gets around the overfitting problem

Problems:

- Requires brute-force spectral decomposition of many matrices
- Easy to code but slow to execute
- Cannot go much beyond dimension $p = 1000$ computationally

To get an *analytical* solution: need Random Matrix Theory

Outline

Limiting Spectral Distribution

Limiting Spectral Distribution

Assumption 3.1

- *p and n go to infinity with $p/n \to c \in (0,1)$ 'concentration ratio'*

Limiting Spectral Distribution

Assumption 3.1

- *p and n go to infinity with* $p/n \to c \in (0,1)$ *'concentration ratio'*
- *population eigenvalues are* $\tau_{n,1}, \ldots, \tau_{n,p}$

Limiting Spectral Distribution

Assumption 3.1

- *p and n go to infinity with* $p/n \to c \in (0, 1)$ *'concentration ratio'*
- *population eigenvalues are* $\tau_{n,1}, \ldots, \tau_{n,p}$
- *population spectral c.d.f. is* $H_n(x) := p^{-1} \sum_{i=1}^{p} \mathbf{1}_{\{x \geq \tau_i\}}$

Limiting Spectral Distribution

Assumption 3.1

- *p and n go to infinity with* $p/n \to c \in (0,1)$ *'concentration ratio'*
- *population eigenvalues are* $\tau_{n,1}, \ldots, \tau_{n,p}$
- *population spectral c.d.f. is* $H_n(x) := p^{-1} \sum_{i=1}^{p} \mathbf{1}_{\{x \geq \tau_i\}}$
- H_n *converges to some limit* H

Limiting Spectral Distribution

Assumption 3.1

- *p and n go to infinity with $p/n \to c \in (0,1)$ 'concentration ratio'*
- *population eigenvalues are $\tau_{n,1}, \ldots, \tau_{n,p}$*
- *population spectral c.d.f. is $H_n(x) := p^{-1} \sum_{i=1}^{p} \mathbf{1}_{\{x \geq \tau_i\}}$*
- *H_n converges to some limit H*

Remark 3.1

This is *not* the spiked model of Johnstone (2001, AoS), which assumes that, apart from a finite number r of 'spikes', the $p - r$ population eigenvalues in the 'bulk' are equal to one another. By contrast, we can handle the general case with any shape(s) of bulk(s).

Limiting Spectral Distribution

Assumption 3.1

- *p and n go to infinity with $p/n \to c \in (0,1)$ 'concentration ratio'*
- *population eigenvalues are $\tau_{n,1}, \dots, \tau_{n,p}$*
- *population spectral c.d.f. is $H_n(x) := p^{-1} \sum_{i=1}^{p} \mathbf{1}_{\{x \geq \tau_i\}}$*
- *H_n converges to some limit H*

Remark 3.1

This is *not* the spiked model of Johnstone (2001, AoS), which assumes that, apart from a finite number r of 'spikes', the $p - r$ population eigenvalues in the 'bulk' are equal to one another. By contrast, we can handle the general case with any shape(s) of bulk(s).

Theorem 1 (Marčenko and Pastur (1967))

There exists a unique $F := F_{c,H}$ such that the sample spectral c.d.f. $F_n(x) := p^{-1} \sum_{i=1}^{p} \mathbf{1}_{\{x \geq \lambda_{n,i}\}}$ converges to $F(x)$.

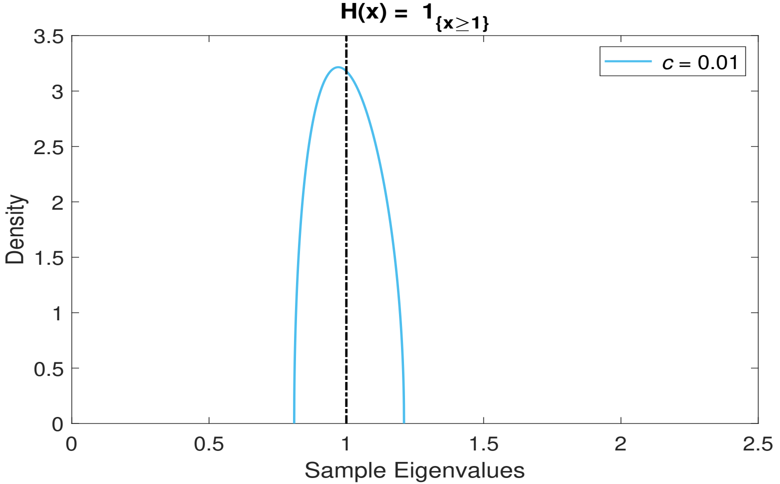
Σ_n = Identity Matrix: Marčenko-Pastur Law

Σ_n = Identity Matrix: Marčenko-Pastur Law

$$\forall x \in [a_-, a_+] \quad f_{c,H}(x) := \frac{\sqrt{(a_+ - x)(x - a_-)}}{2\pi c x} \quad \text{where } a_\pm := \left(1 \pm \sqrt{c}\right)^2$$

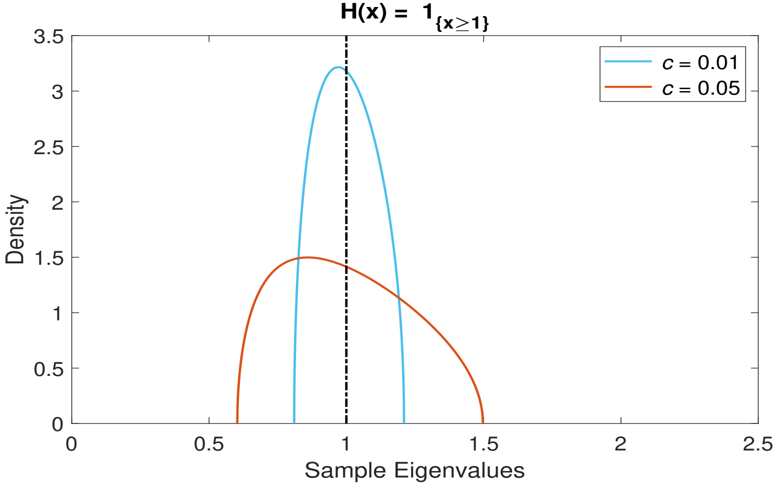
Σ_n = Identity Matrix: Marčenko-Pastur Law

$$\forall x \in [a_-, a_+] \quad f_{c,H}(x) := \frac{\sqrt{(a_+ - x)(x - a_-)}}{2\pi c x} \quad \text{where } a_\pm := \left(1 \pm \sqrt{c}\right)^2$$



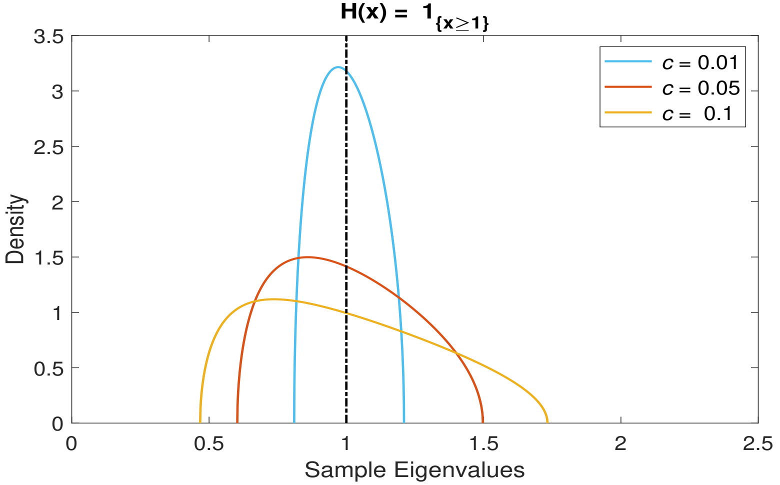
Σ_n = Identity Matrix: Marčenko-Pastur Law

$$\forall x \in [a_-, a_+] \quad f_{c,H}(x) := \frac{\sqrt{(a_+ - x)(x - a_-)}}{2\pi c x} \quad \text{where } a_\pm := \left(1 \pm \sqrt{c}\right)^2$$



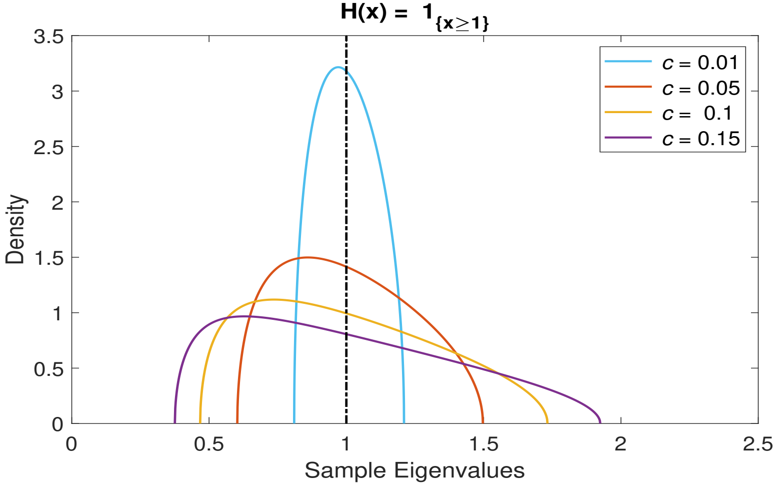
Σ_n = Identity Matrix: Marčenko-Pastur Law

$$\forall x \in [a_-, a_+] \quad f_{c,H}(x) := \frac{\sqrt{(a_+ - x)(x - a_-)}}{2\pi c x} \quad \text{where } a_\pm := \left(1 \pm \sqrt{c}\right)^2$$



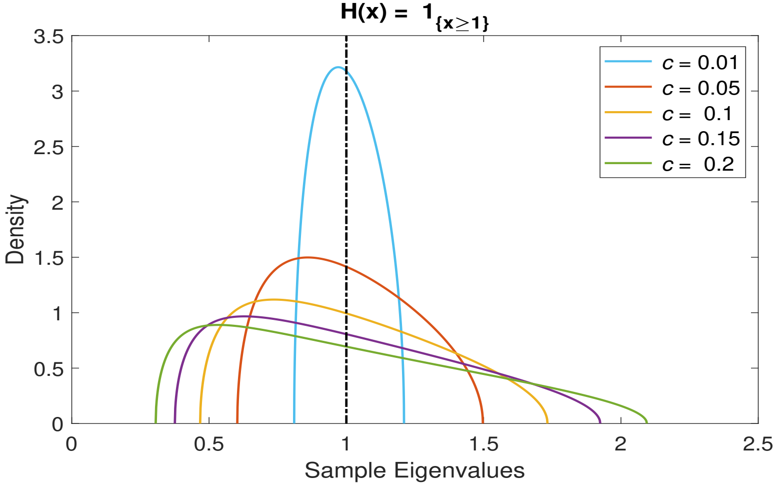
Σ_n = Identity Matrix: Marčenko-Pastur Law

$$\forall x \in [a_-, a_+] \quad f_{c,H}(x) := \frac{\sqrt{(a_+ - x)(x - a_-)}}{2\pi c x} \quad \text{where } a_\pm := \left(1 \pm \sqrt{c}\right)^2$$



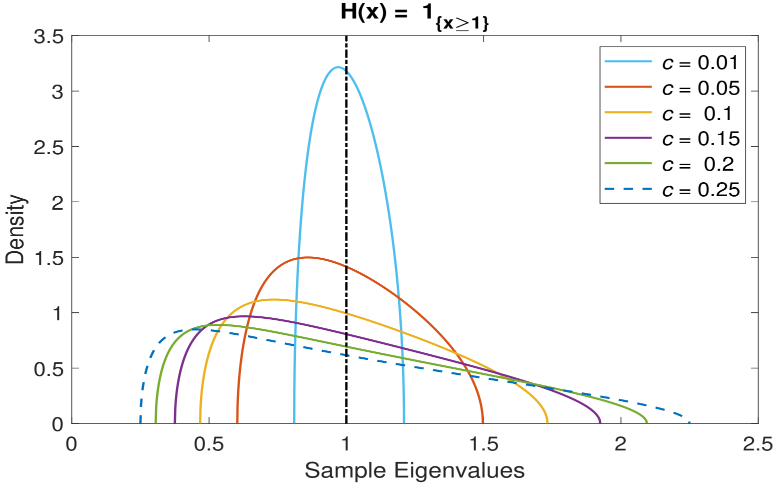
Σ_n = Identity Matrix: Marčenko-Pastur Law

$$\forall x \in [a_-, a_+] \quad f_{c,H}(x) := \frac{\sqrt{(a_+ - x)(x - a_-)}}{2\pi c x} \quad \text{where } a_\pm := \left(1 \pm \sqrt{c}\right)^2$$



Σ_n = Identity Matrix: Marčenko-Pastur Law

$$\forall x \in [a_-, a_+] \quad f_{c,H}(x) := \frac{\sqrt{(a_+ - x)(x - a_-)}}{2\pi c x} \quad \text{where } a_\pm := \left(1 \pm \sqrt{c}\right)^2$$



General Case: True Covariance Matrix $\neq$ Identity

General Case: True Covariance Matrix ≠ Identity

Definition 2 (Stieltjes Transform)

The Stieltjes transform of F is $m_F(z) := \int (\lambda - z)^{-1} dF(\lambda)$
for $z \in \mathbb{C}^+$: complex numbers with imaginary part > 0.

General Case: True Covariance Matrix ≠ Identity

Definition 2 (Stieltjes Transform)

The Stieltjes transform of F is $\quad m_F(z) := \int (\lambda - z)^{-1} dF(\lambda)$
for $z \in \mathbb{C}^+$: complex numbers with imaginary part > 0.

Theorem 3 (Silverstein and Bai (1995); Silverstein (1995))

$m \equiv m_F(z)$ is the unique solution in $\mathbb{C}^+$ to

$$m = \int_{-\infty}^{+\infty} \frac{dH(\tau)}{\tau\,[1 - c - c\,z\,m] - z}$$

General Case: True Covariance Matrix ≠ Identity

Definition 2 (Stieltjes Transform)

The Stieltjes transform of F is $\quad m_F(z) := \int (\lambda - z)^{-1} dF(\lambda)$
for $z \in \mathbb{C}^+$: complex numbers with imaginary part > 0.

Theorem 3 (Silverstein and Bai (1995); Silverstein (1995))

$m \equiv m_F(z)$ is the unique solution in $\mathbb{C}^+$ to

$$m = \int_{-\infty}^{+\infty} \frac{dH(\tau)}{\tau\,[1 - c - c\,z\,m] - z}$$

Theorem 4 (Silverstein and Choi (1995))

m_F admits a continuous extension to the real line $\breve{m}_F(x) := \lim_{z \in \mathbb{C}^+ \to x} m_F(z)$, and the sample spectral density is $f(x) := F'(x) = \pi^{-1} \mathrm{Im}[\breve{m}_F(x)]$.

General Case: True Covariance Matrix ≠ Identity

Definition 2 (Stieltjes Transform)

The Stieltjes transform of F is $m_F(z) := \int (\lambda - z)^{-1} dF(\lambda)$
for $z \in \mathbb{C}^+$: complex numbers with imaginary part > 0.

Theorem 3 (Silverstein and Bai (1995); Silverstein (1995))

$m \equiv m_F(z)$ is the unique solution in $\mathbb{C}^+$ to

$$m = \int_{-\infty}^{+\infty} \frac{dH(\tau)}{\tau\left[1 - c - c\,z\,m\right] - z}$$

Theorem 4 (Silverstein and Choi (1995))

m_F admits a continuous extension to the real line $\breve{m}_F(x) := \lim_{z \in \mathbb{C}^+ \to x} m_F(z)$, and the sample spectral density is $f(x) := F'(x) = \pi^{-1} \mathrm{Im}[\breve{m}_F(x)]$.

Integrate f, and this is how you go from (c, H) to $F = F_{c,H}$.

A Conjecture on the Inverse Problem when $F = F_{c,H}$

A Conjecture on the Inverse Problem when $F = F_{c,H}$

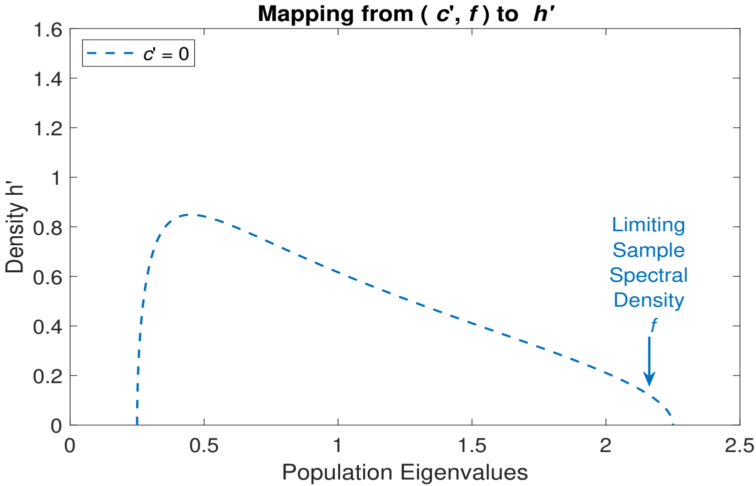
Conjecture 3.1

For every $c' \leq c$, there exists a c.d.f. H' such that $F_{c',H'} = F$.

A Conjecture on the Inverse Problem when $F = F_{c,H}$

Conjecture 3.1

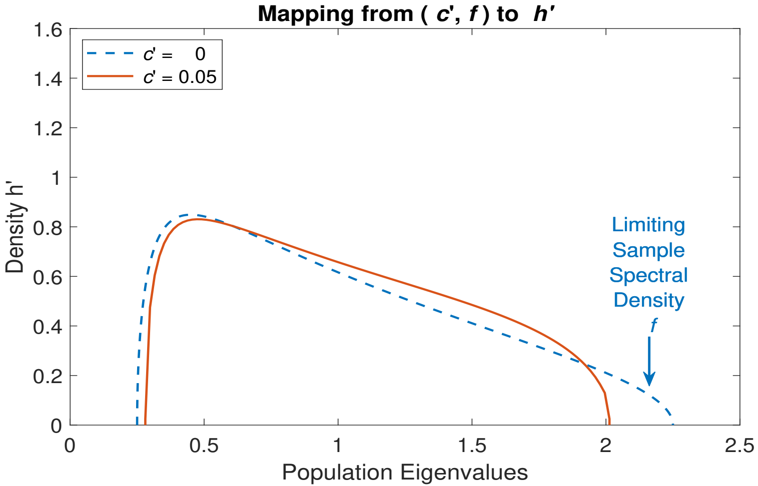
For every $c' \leq c$, there exists a c.d.f. H' such that $F_{c',H'} = F$.



A Conjecture on the Inverse Problem when $F = F_{c,H}$

Conjecture 3.1

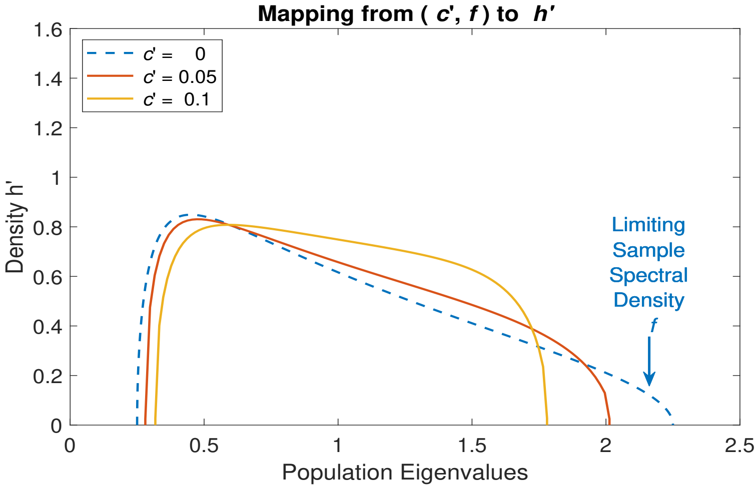
For every $c' \leq c$, there exists a c.d.f. H' such that $F_{c',H'} = F$.



A Conjecture on the Inverse Problem when $F = F_{c,H}$

Conjecture 3.1

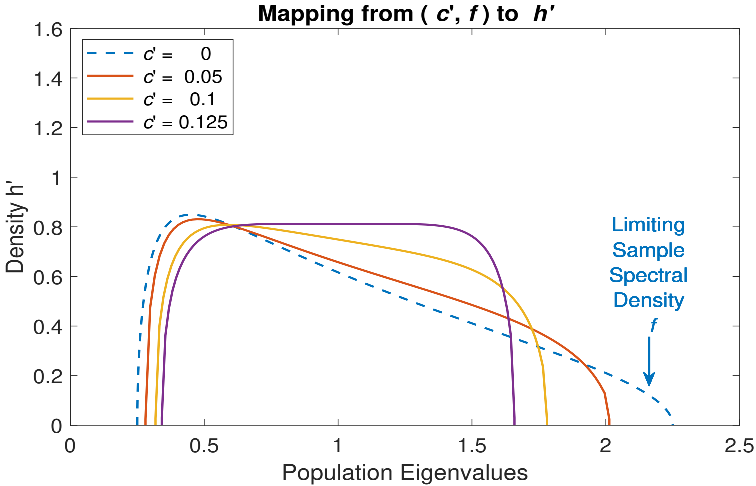
For every $c' \leq c$, there exists a c.d.f. H' such that $F_{c',H'} = F$.



A Conjecture on the Inverse Problem when $F = F_{c,H}$

Conjecture 3.1

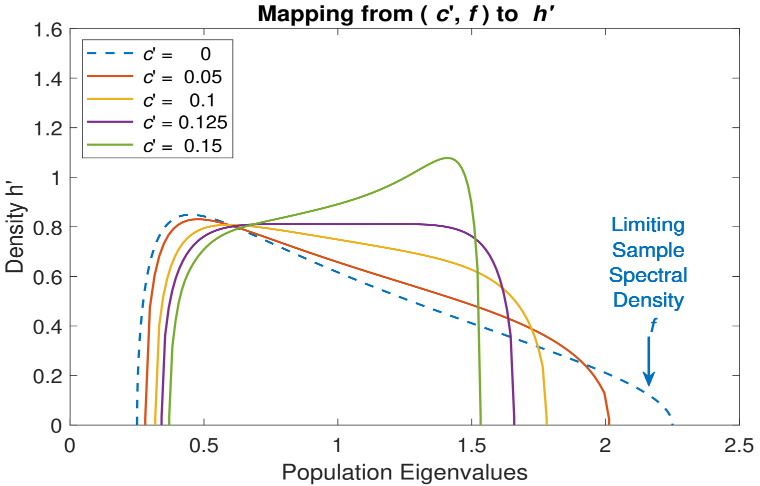
For every $c' \leq c$, there exists a c.d.f. H' such that $F_{c',H'} = F$.



A Conjecture on the Inverse Problem when $F = F_{c,H}$

Conjecture 3.1

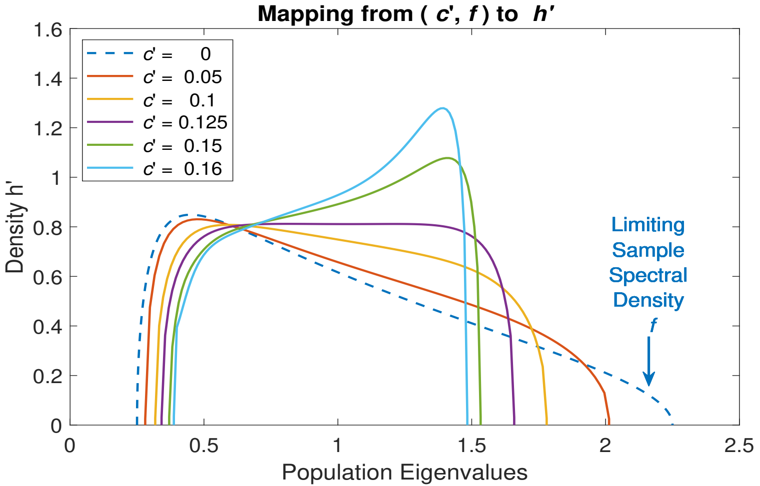
For every $c' \leq c$, there exists a c.d.f. H' such that $F_{c',H'} = F$.



A Conjecture on the Inverse Problem when $F = F_{c,H}$

Conjecture 3.1

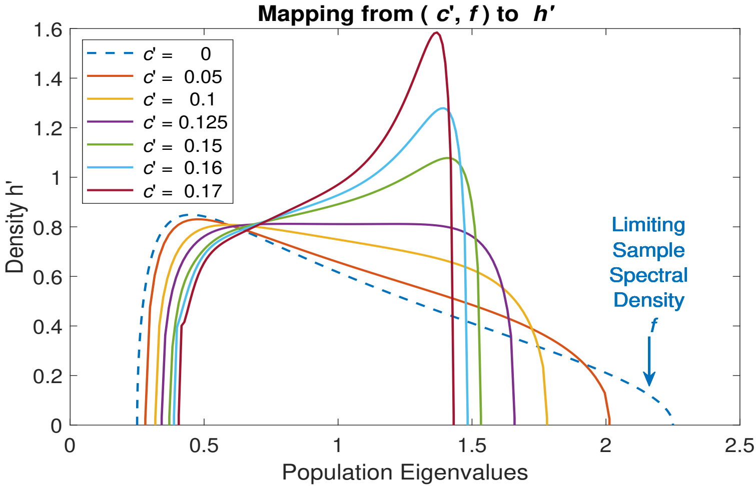
For every $c' \leq c$, there exists a c.d.f. H' such that $F_{c',H'} = F$.



A Conjecture on the Inverse Problem when $F = F_{c,H}$

Conjecture 3.1

For every $c' \leq c$, there exists a c.d.f. H' such that $F_{c',H'} = F$.



The Real Part of the Stieltjes Transform

The Real Part of the Stieltjes Transform

- $\pi^{-1}\text{Im}\,[\breve{m}_F(x)] = f(x)$: the limiting sample spectral density

The Real Part of the Stieltjes Transform

- $\pi^{-1}\mathsf{Im}\,[\breve{m}_F(x)] = f(x)$: the limiting sample spectral density
- $\pi^{-1}\mathsf{Re}\,[\breve{m}_F(x)] = \mathcal{H}_f(x)$: its Hilbert transform

The Real Part of the Stieltjes Transform

- $\pi^{-1}\mathsf{Im}\,[\breve{m}_F(x)] = f(x)$: the limiting sample spectral density
- $\pi^{-1}\mathsf{Re}\,[\breve{m}_F(x)] = \mathcal{H}_f(x)$: its Hilbert transform
- Krantz (2009)

> *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.*

The Real Part of the Stieltjes Transform

- $\pi^{-1}\mathrm{Im}\,[\breve{m}_F(x)] = f(x)$: the limiting sample spectral density
- $\pi^{-1}\mathrm{Re}\,[\breve{m}_F(x)] = \mathcal{H}_f(x)$: its Hilbert transform
- Krantz (2009)

> *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.*

- Hilbert transform of some p.d.f. g: $\mathcal{H}_g(x) := \pi^{-1} PV \int (t-x)^{-1} g(t)dt$

The Real Part of the Stieltjes Transform

- $\pi^{-1}\text{Im}\,[\breve{m}_F(x)] = f(x)$: the limiting sample spectral density
- $\pi^{-1}\text{Re}\,[\breve{m}_F(x)] = \mathcal{H}_f(x)$: its Hilbert transform
- Krantz (2009)

  > *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.*

- Hilbert transform of some p.d.f. g: $\mathcal{H}_g(x) := \pi^{-1} PV \int (t-x)^{-1} g(t) dt$
- Cauchy Principal Value:

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

The Real Part of the Stieltjes Transform

- $\pi^{-1}\mathrm{Im}\,[\breve{m}_F(x)] = f(x)$: the limiting sample spectral density
- $\pi^{-1}\mathrm{Re}\,[\breve{m}_F(x)] = \mathcal{H}_f(x)$: its Hilbert transform
- Krantz (2009)

  > *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.*

- Hilbert transform of some p.d.f. g: $\mathcal{H}_g(x) := \pi^{-1} PV \int (t-x)^{-1} g(t)dt$
- Cauchy Principal Value:

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

- Highly positive just below the center of mass of the density g

The Real Part of the Stieltjes Transform

- $\pi^{-1}\text{Im}\,[\breve{m}_F(x)] = f(x)$: the limiting sample spectral density
- $\pi^{-1}\text{Re}\,[\breve{m}_F(x)] = \mathcal{H}_f(x)$: its Hilbert transform
- Krantz (2009)

  > *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.*

- Hilbert transform of some p.d.f. g: $\mathcal{H}_g(x) := \pi^{-1} PV \int (t-x)^{-1} g(t)dt$
- Cauchy Principal Value:

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

- Highly positive just below the center of mass of the density g
- Highly negative just above the center of mass of the density g

The Real Part of the Stieltjes Transform

- $\pi^{-1}\mathrm{Im}\,[\breve{m}_F(x)] = f(x)$: the limiting sample spectral density
- $\pi^{-1}\mathrm{Re}\,[\breve{m}_F(x)] = \mathcal{H}_f(x)$: its Hilbert transform
- Krantz (2009)

  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.

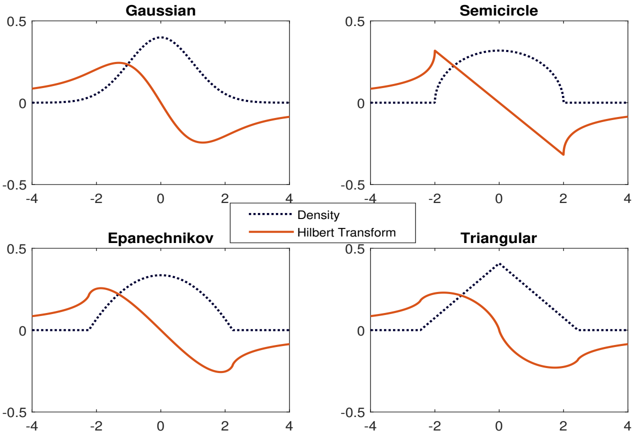
- Hilbert transform of some p.d.f. g: $\mathcal{H}_g(x) := \pi^{-1} PV \int (t-x)^{-1} g(t) dt$
- Cauchy Principal Value:

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

- Highly positive just below the center of mass of the density g
- Highly negative just above the center of mass of the density g
- Fades to zero away from center of mass

Four Examples of Hilbert Transforms

Four Examples of Hilbert Transforms



Works like a local attraction force

Ledoit and Péché (2011)

Ledoit and Péché (2011)

- Finite sample analysis $\Longrightarrow$ estimate the covariance matrix by: (1) keeping the sample eigenvectors $(u_{n,i})_{i=1,\dots,p}$; and (2) replacing the sample eigenvalues $\lambda_{n,i} = u'_{n,i} S_n u_{n,i}$ with $\delta^*_{n,i} = u'_{n,i} \Sigma_n u_{n,i}$

Ledoit and Péché (2011)

- Finite sample analysis $\Longrightarrow$ estimate the covariance matrix by: (1) keeping the sample eigenvectors $(u_{n,i})_{i=1,\dots,p}$; and (2) replacing the sample eigenvalues $\lambda_{n,i} = u'_{n,i} S_n u_{n,i}$ with $\delta^*_{n,i} = u'_{n,i} \Sigma_n u_{n,i}$
- Ledoit-Péché show that under large-dimensional asymptotics:

Ledoit and Péché (2011)

- Finite sample analysis $\Longrightarrow$ estimate the covariance matrix by: (1) keeping the sample eigenvectors $(u_{n,i})_{i=1,\dots,p}$; and (2) replacing the sample eigenvalues $\lambda_{n,i} = u'_{n,i} S_n u_{n,i}$ with $\delta^*_{n,i} = u'_{n,i} \Sigma_n u_{n,i}$
- Ledoit-Péché show that under large-dimensional asymptotics:

$$u'_{n,i} \Sigma_n u_{n,i} \approx \frac{\lambda_{n,i}}{\left[\pi c \lambda_{n,i} f(\lambda_{n,i})\right]^2 + \left[1 - c - \pi c \lambda_{n,i} \mathcal{H}_f(\lambda_{n,i})\right]^2}$$

Ledoit and Péché (2011)

- Finite sample analysis $\Longrightarrow$ estimate the covariance matrix by: (1) keeping the sample eigenvectors $(u_{n,i})_{i=1,\dots,p}$; and (2) replacing the sample eigenvalues $\lambda_{n,i} = u'_{n,i} S_n u_{n,i}$ with $\delta^*_{n,i} = u'_{n,i} \Sigma_n u_{n,i}$
- Ledoit-Péché show that under large-dimensional asymptotics:

$$u'_{n,i} \Sigma_n u_{n,i} \approx \frac{\lambda_{n,i}}{\left[\pi c \lambda_{n,i} f(\lambda_{n,i})\right]^2 + \left[1 - c - \pi c \lambda_{n,i} \mathcal{H}_f(\lambda_{n,i})\right]^2}$$

where $f := f_{c,H}$ is the limiting spectral density

Ledoit and Péché (2011)

- Finite sample analysis $\Longrightarrow$ estimate the covariance matrix by: (1) keeping the sample eigenvectors $(u_{n,i})_{i=1,\dots,p}$; and (2) replacing the sample eigenvalues $\lambda_{n,i} = u'_{n,i} S_n u_{n,i}$ with $\delta^*_{n,i} = u'_{n,i} \Sigma_n u_{n,i}$
- Ledoit-Péché show that under large-dimensional asymptotics:

$$u'_{n,i} \Sigma_n u_{n,i} \approx \frac{\lambda_{n,i}}{\left[\pi c \lambda_{n,i} f(\lambda_{n,i})\right]^2 + \left[1 - c - \pi c \lambda_{n,i} \mathcal{H}_f(\lambda_{n,i})\right]^2}$$

  where $f := f_{c,H}$ is the limiting spectral density
- This is an *oracle* formula because f and $\mathcal{H}_f$ are unknown

Ledoit and Péché (2011)

- Finite sample analysis $\Longrightarrow$ estimate the covariance matrix by: (1) keeping the sample eigenvectors $(u_{n,i})_{i=1,\dots,p}$; and (2) replacing the sample eigenvalues $\lambda_{n,i} = u'_{n,i} S_n u_{n,i}$ with $\delta^*_{n,i} = u'_{n,i} \Sigma_n u_{n,i}$
- Ledoit-Péché show that under large-dimensional asymptotics:

$$u'_{n,i} \Sigma_n u_{n,i} \approx \frac{\lambda_{n,i}}{\left[\pi c \lambda_{n,i} f(\lambda_{n,i})\right]^2 + \left[1 - c - \pi c \lambda_{n,i} \mathcal{H}_f(\lambda_{n,i})\right]^2}$$

  where $f := f_{c,H}$ is the limiting spectral density
- This is an *oracle* formula because f and $\mathcal{H}_f$ are unknown
- Results in local attraction: any sample eigenvalue moves toward the mass center closest to it

Ledoit and Péché (2011)

- Finite sample analysis $\Longrightarrow$ estimate the covariance matrix by: (1) keeping the sample eigenvectors $(u_{n,i})_{i=1,\dots,p}$; and (2) replacing the sample eigenvalues $\lambda_{n,i} = u'_{n,i} S_n u_{n,i}$ with $\delta^*_{n,i} = u'_{n,i} \Sigma_n u_{n,i}$
- Ledoit-Péché show that under large-dimensional asymptotics:

$$u'_{n,i} \Sigma_n u_{n,i} \approx \frac{\lambda_{n,i}}{\left[\pi c \lambda_{n,i} f(\lambda_{n,i})\right]^2 + \left[1 - c - \pi c \lambda_{n,i} \mathcal{H}_f(\lambda_{n,i})\right]^2}$$

  where $f := f_{c,H}$ is the limiting spectral density
- This is an *oracle* formula because f and $\mathcal{H}_f$ are unknown
- Results in local attraction: any sample eigenvalue moves toward the mass center closest to it
- Different from Ledoit and Wolf (2004) linear shrinkage, where all eigenvalues move to the same *global* center of mass

Ledoit and Péché (2011)

- Finite sample analysis $\Longrightarrow$ estimate the covariance matrix by: (1) keeping the sample eigenvectors $(u_{n,i})_{i=1,\ldots,p}$; and (2) replacing the sample eigenvalues $\lambda_{n,i} = u'_{n,i} S_n u_{n,i}$ with $\delta^*_{n,i} = u'_{n,i} \Sigma_n u_{n,i}$
- Ledoit-Péché show that under large-dimensional asymptotics:

$$u'_{n,i}\Sigma_n u_{n,i} \approx \frac{\lambda_{n,i}}{\left[\pi c \lambda_{n,i} f(\lambda_{n,i})\right]^2 + \left[1 - c - \pi c \lambda_{n,i} \mathcal{H}_f(\lambda_{n,i})\right]^2}$$

  where $f := f_{c,H}$ is the limiting spectral density
- This is an *oracle* formula because f and $\mathcal{H}_f$ are unknown
- Results in local attraction: any sample eigenvalue moves toward the mass center closest to it
- Different from Ledoit and Wolf (2004) linear shrinkage, where all eigenvalues move to the same *global* center of mass
- Need to shrink **within-clusters**, not so much **between-clusters**

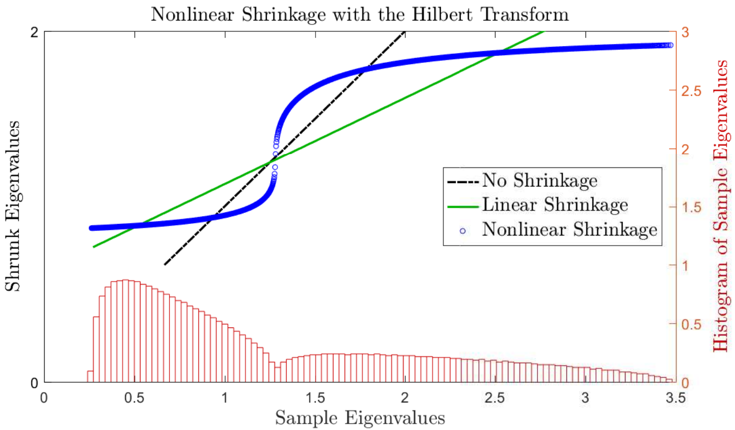
Nonlinear Shrinkage Is Local Shrinkage

Nonlinear Shrinkage Is Local Shrinkage

Σ_n: $2,500$ eigenvalues equal to 0.8 and $1,500$ equal to 2; $n = 18,000$

Nonlinear Shrinkage Is Local Shrinkage

Σ_n: 2,500 eigenvalues equal to 0.8 and 1,500 equal to 2; $n = 18,000$



How to Estimate f and its Hilbert Transform?

How to Estimate f and its Hilbert Transform?

- Indirect approach: go through population spectral c.d.f. H

How to Estimate f and its Hilbert Transform?

- Indirect approach: go through population spectral c.d.f. H
- QuEST function (Ledoit and Wolf, 2015) maps $(c, H) \mapsto F_{c,H}$

How to Estimate f and its Hilbert Transform?

- Indirect approach: go through population spectral c.d.f. H
- QuEST function (Ledoit and Wolf, 2015) maps $(c, H) \mapsto F_{c,H}$
- Step 1: given observed $F_n(x) := p^{-1} \sum_{i=1}^{p} \mathbf{1}_{\{x \geq \lambda_{n,i}\}}$, find $\widehat{H}_n$ that provides the best match

How to Estimate f and its Hilbert Transform?

- Indirect approach: go through population spectral c.d.f. H
- QuEST function (Ledoit and Wolf, 2015) maps $(c, H) \mapsto F_{c,H}$
- Step 1: given observed $F_n(x) := p^{-1} \sum_{i=1}^{p} \mathbf{1}_{\{x \geq \lambda_{n,i}\}}$, find $\widehat{H}_n$ that provides the best match
- Step 2: Given c and $\widehat{H}_n$, compute Stieltjes transform of $F_{c,\widehat{H}_n}$

How to Estimate f and its Hilbert Transform?

- Indirect approach: go through population spectral c.d.f. H
- QuEST function (Ledoit and Wolf, 2015) maps $(c, H) \mapsto F_{c,H}$
- Step 1: given observed $F_n(x) := p^{-1} \sum_{i=1}^{p} \mathbf{1}_{\{x \geq \lambda_{n,i}\}}$, find $\widehat{H}_n$ that provides the best match
- Step 2: Given c and $\widehat{H}_n$, compute Stieltjes transform of $F_{c,\widehat{H}_n}$
- **Problem:** Step 1 solves numerically a high-dimensional constrained nonlinear minimization problem $\longrightarrow$ slow, and hard to scale above dimension $p = 1,000$

How to Estimate f and its Hilbert Transform?

- Indirect approach: go through population spectral c.d.f. H
- QuEST function (Ledoit and Wolf, 2015) maps $(c, H) \mapsto F_{c,H}$
- Step 1: given observed $F_n(x) := p^{-1} \sum_{i=1}^{p} \mathbf{1}_{\{x \geq \lambda_{n,i}\}}$, find $\widehat{H}_n$ that provides the best match
- Step 2: Given c and $\widehat{H}_n$, compute Stieltjes transform of $F_{c,\widehat{H}_n}$
- **Problem:** Step 1 solves numerically a high-dimensional constrained nonlinear minimization problem $\longrightarrow$ slow, and hard to scale above dimension $p = 1,000$
- Also: population spectrum is a *nuisance parameter* with no direct bearing on the outcome

How to Estimate f and its Hilbert Transform?

- Indirect approach: go through population spectral c.d.f. H
- QuEST function (Ledoit and Wolf, 2015) maps $(c, H) \mapsto F_{c,H}$
- Step 1: given observed $F_n(x) := p^{-1} \sum_{i=1}^{p} \mathbf{1}_{\{x \geq \lambda_{n,i}\}}$, find $\widehat{H}_n$ that provides the best match
- Step 2: Given c and $\widehat{H}_n$, compute Stieltjes transform of $F_{c,\widehat{H}_n}$
- **Problem:** Step 1 solves numerically a high-dimensional constrained nonlinear minimization problem $\longrightarrow$ slow, and hard to scale above dimension $p = 1,000$
- Also: population spectrum is a *nuisance parameter* with no direct bearing on the outcome

It would be nice to have a direct estimator for f and $\mathcal{H}_f$ that depends only on sample eigenvalues, with fast analytical formula.

Outline

Choice of Kernel

Kernel estimation of limiting sample spectral density was pioneered by Bing-Yi Jing, Guangming Pan, Qi-Man Shao and Wang Zhou (2010, AoS).

Choice of Kernel

Kernel estimation of limiting sample spectral density was pioneered by Bing-Yi Jing, Guangming Pan, Qi-Man Shao and Wang Zhou (2010, AoS).

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 well-known Epanechnikov kernel.

We also prove that it satisfies all the above assumptions.

Choice of Bandwidth

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 $h_n \to 0$.

Jing et al. (2010) used $h_n := n^{-1/3}$, so we keep the same exponent.

Note:

- They actually use a uniform bandwidth $h_{n,i} \equiv n^{-1/3}$
- This results in worse finite-sample performance
- Also fails to respect the scale-equivariant nature of the problem

Kernel Estimators & Feasible Shrinkage Formula

Kernel estimators of f and $\mathcal{H}_f$

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

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

Feasible analytical nonlinear shrinkage estimator of Σ_n

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

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

Closing Thoughts on Kernel Estimation

Closing Thoughts on Kernel Estimation

The 2010 paper by Jing, Pan, Shao and Zhou was entitled "Nonparametric estimate of spectral density functions of sample covariance matrices: A first step".

Closing Thoughts on Kernel Estimation

The 2010 paper by Jing, Pan, Shao and Zhou was entitled "Nonparametric estimate of spectral density functions of sample covariance matrices: A first step".

At the narrowest level, we do "A second step" by:

- moving from fixed to proportional bandwidth,
- generalizing their results to obtain a nonparametric estimate of the Hilbert transform of the spectral density of the sample covariance matrix.

Closing Thoughts on Kernel Estimation

The 2010 paper by Jing, Pan, Shao and Zhou was entitled "Nonparametric estimate of spectral density functions of sample covariance matrices: A first step".

At the narrowest level, we do "A second step" by:

- moving from fixed to proportional bandwidth,
- generalizing their results to obtain a nonparametric estimate of the Hilbert transform of the spectral density of the sample covariance matrix.

But our main contribution is to harness the technique to make headway on the general problem of estimating the covariance matrix.

Closing Thoughts on Kernel Estimation

The 2010 paper by Jing, Pan, Shao and Zhou was entitled "Nonparametric estimate of spectral density functions of sample covariance matrices: A first step".

At the narrowest level, we do "A second step" by:

- moving from fixed to proportional bandwidth,
- generalizing their results to obtain a nonparametric estimate of the Hilbert transform of the spectral density of the sample covariance matrix.

But our main contribution is to harness the technique to make headway on the general problem of estimating the covariance matrix.

The hard work of connecting the pipes (mathematically speaking) happens essentially 'behind the scene', and it owes much debt to foundational results first laid out in Ledoit and Wolf (2012, AoS).

Outline

Executive Summary

Performance of analytical nonlinear shrinkage:

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

Speed of analytical nonlinear shrinkage:

- Basically as fast as linear shrinkage
- Much faster than QuEST
- Much faster than NERCOME

$\Longrightarrow$ Get the best of both worlds!

Main Performance Measure

Percentage Relative Improvement in Average Loss (PRIAL):

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

By construction:

- The sample covariance matrix S_n has $\mathrm{PRIAL}_n^{\mathrm{MV}}\big(S_n\big) = 0\%$
- The FSOPT **'Gold Standard'** has $\mathrm{PRIAL}_n^{\mathrm{MV}}\big(S_n^*\big) = 100\%$

Note:

- Negative PRIAL values are possible

Baseline Scenario

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

- Dimension $p = 200$
- Sample size $n = 600$
- Concentration ratio $\widehat{c_n} = 1/3$
- 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

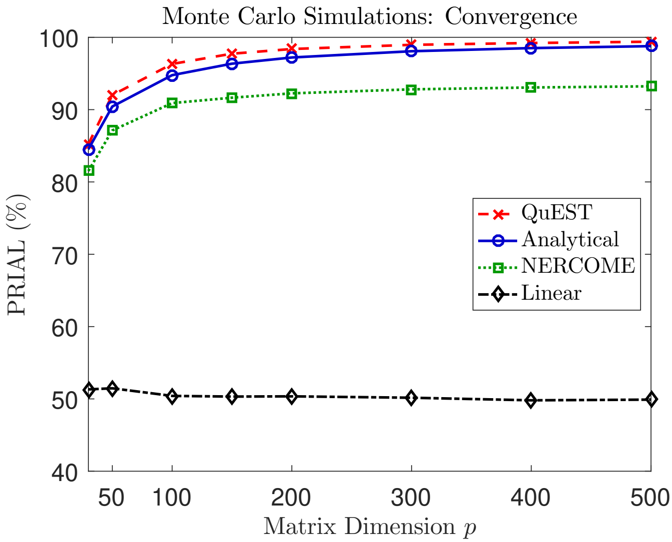
Estimator	Sample	Linear	Analytical	QuEST	NERCOME	FSOPT
∅ Loss	2.71	2.10	1.52	1.50	1.58	1.48
PRIAL	0%	50%	97%	98%	92%	100%
Time (ms)	1	3	4	2,233	2,990	3

Note:

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

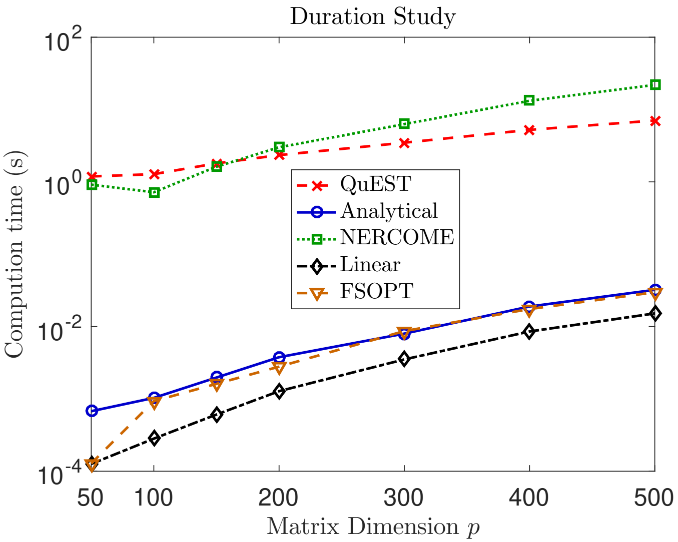
Large-Dimensional Asymptotics

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



Speed

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:

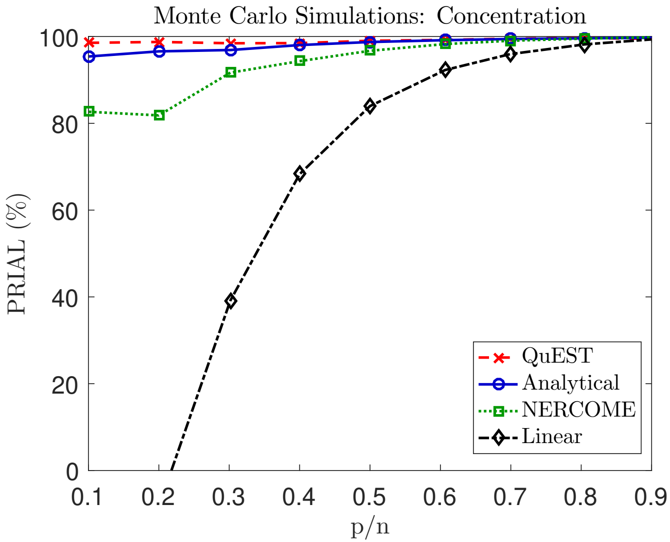
- $p = 10,000$
- $n = 30,000$

QuEST and NERCOME are no longer computationally feasible.

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

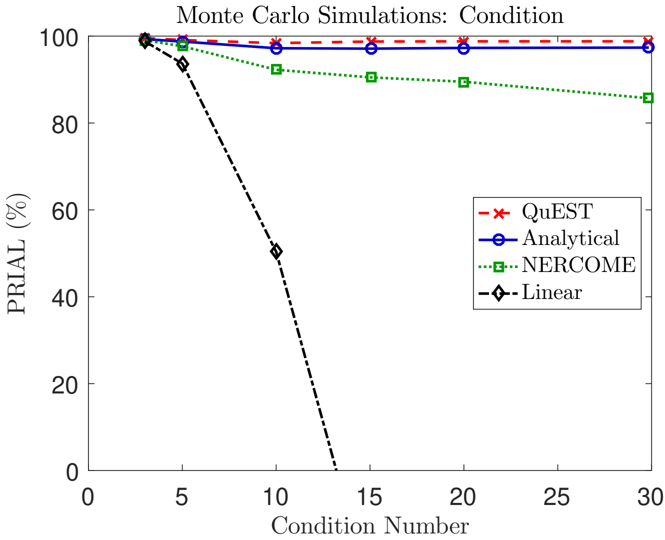
Concentration Ratio

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



Condition Number

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



Non-Normality

Vary the distribution of the variates:

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

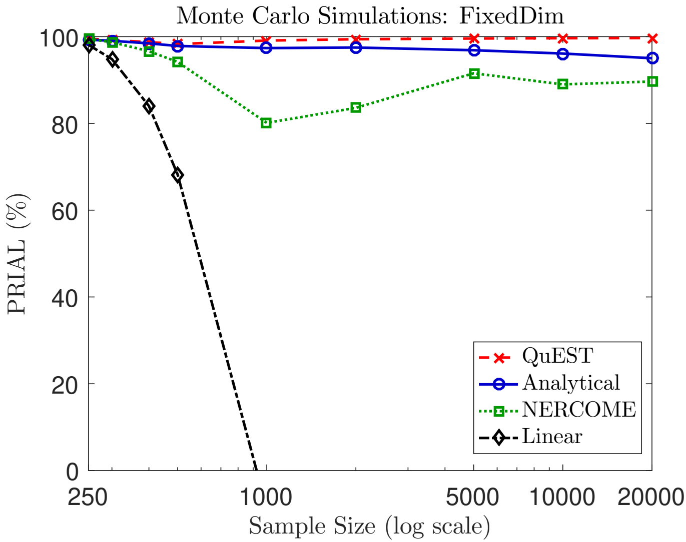
Shape of Distribution of Population Eigenvalues

Use a shifted and stretched Beta distribution with support [1,10]:

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%

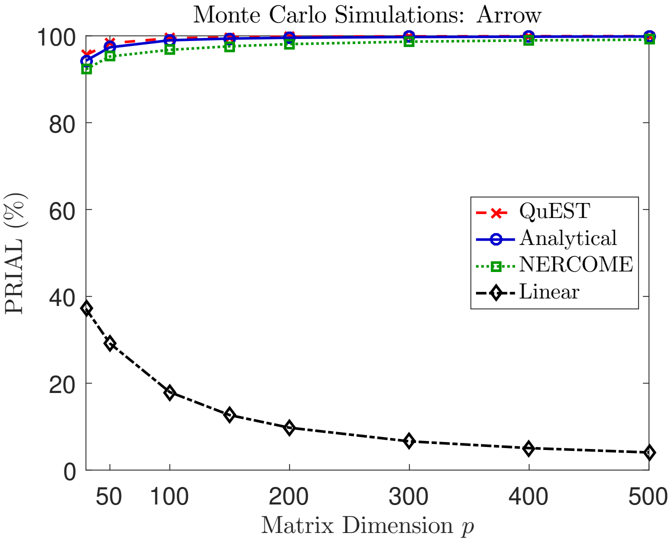
Fixed-Dimensional Asymptotics

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



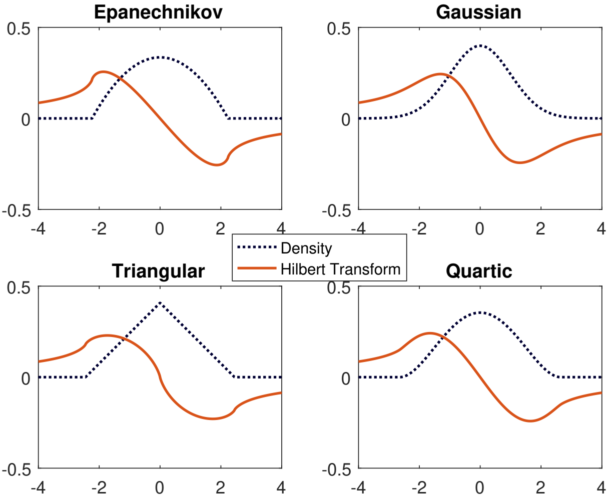
Arrow Model

Let $\tau_{n,p} := 1 + 0.5(p-1)$ and remaining bulk from s&s Beta(5,2):



Robustness Check: Choice of Kernel

Consider alternative choices of the kernel:



Robustness Check: Choice of Kernel

Just as good:

- Semi-circle kernel
- Triangular kernel

No good:

- Gaussian kernel (extremely slow)
- Quartic kernel (numerical issues)

Robustness Check: Multiplier and Exponent

Consider a base-rate bandwidth of the form $h_n := Kn^{-\alpha}$ with

- $K \in \{0.5, 1, 2\}$
- $\alpha \in \{0.2, 0.25, 0.3, 1/3, 0.35\}$

Finding:

- Our initial choices $K = 1$ and $\alpha = 1/3$ cannot be bettered

Additional finding:

- Using a uniform bandwidth $h_{n,i} \equiv \bar{\lambda}_n h_n$ instead of our variable bandwidth $h_{n,i} := \lambda_{n,i} h_n$ reduces performance

Outline

Data & Portfolio Rules

Stocks:

- Download daily return data from CRSP
- Period: 01/01/1973–12/31/2017

Updating:

- 21 consecutive trading days constitute one 'month'
- Update portfolios on 'monthly' basis

Out-of-sample period:

- Start out-of-sample investing on 01/16/1978
- This results in 10,080 daily returns (over 480 'months')

Data & Portfolio Rules

Portfolio sizes:

- We consider $p \in \{100, 500, 1000\}$

Portfolio constituents:

- Select new constituents at the beginning of each month
- If there are pairs of highly correlated stocks ($r > 0.95$), kick out the stock with lower market capitalization
- Find the p largest remaining stocks that have
  (i) a nearly complete 1260-day return history
  (ii) a complete 21-day return future

Estimation:

- Use the previous $n = 1260$ days to estimate the covariance matrix

Global Minimum Variance Portfolio

Problem Formulation:

$$\min_{w} w'\Sigma w$$
$$\text{subject to} \quad w'\mathbf{1} = 1$$

(where **1** is a conformable vector of ones)

Analytical Solution:

$$w^* = \frac{\Sigma^{-1}\mathbf{1}}{\mathbf{1}'\Sigma^{-1}\mathbf{1}}$$

Feasible Solution:

$$\hat{w} := \frac{\hat{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}'\hat{\Sigma}^{-1}\mathbf{1}}$$

Performance Measures

All measures are based on the 10,080 out-of-sample returns and are annualized for convenience.

Performance measures:

- **AV:** Average
- **SD:** Standard deviation (of main interest)
- **IR:** Information ratio, defined as **AV/SD**

Assessing statistical significance:

- We test for outperformance of NonLin over Spiked in terms of SD
- Test is based on Ledoit and Wolf (2011, WM)

Performance Measures

	$p = 100$			$p = 500$			$p = 1000$		
	AV	SD	IR	AV	SD	IR	AV	SD	IR
Identity	12.82	17.40	0.74	13.86	16.83	0.82	14.36	16.85	0.85
Sample	11.94	11.88	1.01	11.89	9.45	1.26	11.83	11.44	1.03
Linear	12.01	11.81	1.02	12.02	9.06	1.33	12.26	8.27	1.48
Spiked	11.92	11.88	1.00	12.27	8.86	1.38	12.51	7.58	1.65
NonLin	11.94	**11.74*****	1.02	11.91	**8.63*****	1.38	12.28	**7.45*****	1.65

Note: In the columns labeled “SD”, the best numbers are in **blue**.

Outline

Conclusion

Conclusion

We view FSOPT (replacing sample eigenvalues with $u_{n,i}' \Sigma_n u_{n,i}$) as the **'Gold Standard'** for covariance matrix estimation because it is the most general solution:

Conclusion

We view FSOPT (replacing sample eigenvalues with $u'_{n,i}\Sigma_n u_{n,i}$) as the **'Gold Standard'** for covariance matrix estimation because it is the most general solution:

- the orientation of the population eigenvectors can be anything,
- the distribution of the population eigenvalues can be anything,
- the shape of the shrinkage function can be anything.

Conclusion

We view FSOPT (replacing sample eigenvalues with $u'_{n,i}\Sigma_n u_{n,i}$) as the **'Gold Standard'** for covariance matrix estimation because it is the most general solution:

- the orientation of the population eigenvectors can be anything,
- the distribution of the population eigenvalues can be anything,
- the shape of the shrinkage function can be anything.

Our estimator is the first analytical formula that attains FSOPT performance under large-dimensional asymptotics. The advantages of being analytical are:

Conclusion

We view FSOPT (replacing sample eigenvalues with $u'_{n,i}\Sigma_n u_{n,i}$) as the **'Gold Standard'** for covariance matrix estimation because it is the most general solution:

- the orientation of the population eigenvectors can be anything,
- the distribution of the population eigenvalues can be anything,
- the shape of the shrinkage function can be anything.

Our estimator is the first analytical formula that attains FSOPT performance under large-dimensional asymptotics. The advantages of being analytical are:

- it is easily understandable and teachable,
- it is fast and scalable up to $10,000$ variables,
- it can be programmed *inside* a further numerical scheme.

Conclusion

We view FSOPT (replacing sample eigenvalues with $u'_{n,i}\Sigma_n u_{n,i}$) as the **'Gold Standard'** for covariance matrix estimation because it is the most general solution:

- the orientation of the population eigenvectors can be anything,
- the distribution of the population eigenvalues can be anything,
- the shape of the shrinkage function can be anything.

Our estimator is the first analytical formula that attains FSOPT performance under large-dimensional asymptotics. The advantages of being analytical are:

- it is easily understandable and teachable,
- it is fast and scalable up to $10,000$ variables,
- it can be programmed *inside* a further numerical scheme.

There are many Big Data M.Sc. programs in their infancy, and the first one to offer a course entitled **"Shrinkage for Big Data"** will gain an edge over the competition.

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

Bell, P. and King, S. (2009). Diagonal priors for full covariance speech recognition. In *IEEE Workshop on Automatic Speech Recognition & Understanding, 2009. ASRU 2009*, pages 113–117. IEEE.

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

Elsheikh, A. H., Wheeler, M. F., and Hoteit, I. (2013). An iterative stochastic ensemble method for parameter estimation of subsurface flow models. *Journal of Computational Physics*, 242:696–714.

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

Fritsch, V., Varoquaux, G., Thyreau, B., Poline, J.-B., and Thirion, B. (2012). Detecting outliers in high-dimensional neuroimaging datasets with robust covariance estimators. *Medical Image Analysis*, 16(7):1359–1370.

Guo, S.-M., He, J., Monnier, N., Sun, G., Wohland, T., and Bathe, M. (2012). Bayesian approach to the analysis of fluorescence correlation spectroscopy data II: application to simulated and in vitro data. *Analytical Chemistry*, 84(9):3880–3888.

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

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.

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

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.

Lin, J.-A., Zhu, H. b., Knickmeyer, R., Styner, M., Gilmore, J., and Ibrahim, J. (2012). Projection regression models for multivariate imaging phenotype. *Genetic Epidemiology*, 36(6):631–641.

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

Markon, K. (2010). Modeling psychopathology structure: A symptom-level analysis of axis I and II disorders. *Psychological Medicine*, 40(2):273–288.

Michaelides, P., Apostolellis, P., and Fassois, S. (2011). Vibration-based damage diagnosis in a laboratory cable–stayed bridge model via an rcp–arx model based method. In *Journal of Physics: Conference Series*, volume 305, page 012104. IOP Publishing.

Pyeon, D., Newton, M., Lambert, P., Den Boon, J., Sengupta, S., Marsit, C., Woodworth, C., Connor, J., Haugen, T., Smith, E., Kelsey, K., Turek, L., and Ahlquist, P. (2007). Fundamental differences in cell cycle deregulation in human papillomavirus-positive and human papillomavirus-negative head/neck and cervical cancers. *Cancer Research*, 67(10):4605–4619.

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.

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.

Wei, Z., Huang, J., and Hui, Y. (2011). Adaptive-beamforming-based multiple targets signal separation. In *Signal Processing, Communications and Computing (ICSPCC), 2011 IEEE International Conference on*, pages 1–4. IEEE.