Spectral analysis of high-dimensional time series with applications to the mean-variance frontier

Alexander Aue

Department of Statistics & GGAM

University of California, Davis

aaue@ucdavis.edu

With Haoyang Liu (Finance, Florida State) and Debashis Paul (Statistics, UC Davis)

Research partially supported by NSF grant DMS 1407530

Outline

A. Introduction

- High-dimensional statistics & random matrix theory
- The sample covariance matrix
- Results for the empirical spectral distribution in the i.i.d. case
- Existing literature on the dependent case

Outline

A. Introduction

- High-dimensional statistics & random matrix theory
- The sample covariance matrix
- Results for the empirical spectral distribution in the i.i.d. case
- Existing literature on the dependent case

B. Spectral Theory for Linear Time Series

- Eigenvalue distribution of sample covariance matrix
- Eigenvalue distribution of symmetrized sample autocovariance matrix
- Proof techniques for MA(1) case

Outline

A. Introduction

- High-dimensional statistics & random matrix theory
- The sample covariance matrix
- Results for the empirical spectral distribution in the i.i.d. case
- Existing literature on the dependent case

B. Spectral Theory for Linear Time Series

- Eigenvalue distribution of sample covariance matrix
- Eigenvalue distribution of symmetrized sample autocovariance matrix
- Proof techniques for MA(1) case

C. Estimation of Quadratic Forms for Time Series

- Uses spectral theory results
- Applies to mean-variance frontier estimation in finance
- Uses thresholding and cross-validation approach
- Empirical results

Outline

A. Introduction

- High-dimensional statistics & random matrix theory
- The sample covariance matrix
- Results for the empirical spectral distribution in the i.i.d. case
- Existing literature on the dependent case

B. Spectral Theory for Linear Time Series

- Eigenvalue distribution of sample covariance matrix
- Eigenvalue distribution of symmetrized sample autocovariance matrix
- Proof techniques for MA(1) case

C. Estimation of Quadratic Forms for Time Series

- Uses spectral theory results
- Applies to mean-variance frontier estimation in finance
- Uses thresholding and cross-validation approach
- Empirical results

D. Conclusions

A. Introduction

Random Matrix Theory (RMT)

- *Origins of RMT*
  - Initially used in physics to study quantum phenomena of heavy atoms
  - Energy levels of a system described by eigenvalues of Hamiltonian operator
  - Explicit calculations only possible for low-energy levels but not for high-energy levels
  - *Wigner (1955, 1958):* Energy levels described by eigenvalues of random matrix

Random Matrix Theory (RMT)

- *Origins of RMT*
  - Initially used in physics to study quantum phenomena of heavy atoms
  - Energy levels of a system described by eigenvalues of Hamiltonian operator
  - Explicit calculations only possible for low-energy levels but not for high-energy levels
  - *Wigner (1955, 1958):* Energy levels described by eigenvalues of random matrix

- *Applications of RMT in statistics*
  - Include problems in dimension reduction, hypothesis testing, clustering, regression analysis and covariance estimation
  - Much of the literature covers the behavior of the sample covariance matrix and the
    - ∗ behavior of the bulk spectrum: empirical spectral distribution
    - ∗ behavior of the edge of the spectrum: extreme (largest/smallest) eigenvalues
    - ∗ distribution of spacings of eigenvalues
    - ∗ behavior of eigenvectors
  - Paul & A (2014), review paper

Asymptotic Setting

- *Connecting dimension with sample size*
  - Suppose $\mathbf{X}$ is a $p \times n$ matrix with real- or complex-valued entries and independent columns
  - Specify that $p = p(n)$ and that

$$\lim_{n\to\infty} \frac{p}{n} = \gamma \in (0, \infty) \tag{1}$$

Asymptotic Setting

- *Connecting dimension with sample size*
  - Suppose $\mathbf{X}$ is a $p \times n$ matrix with real- or complex-valued entries and independent columns
  - Specify that $p = p(n)$ and that

$$\lim_{n \to \infty} \frac{p}{n} = \gamma \in (0, \infty) \tag{2}$$

- *Wigner matrices*
  - Used as model for spectra of heavy atoms
  - Here $p = n$ such that $X_{ij} = \overline{X}_{ji}$ (symmetric/Hermitian; diagonal always real-valued)
  - X_{ij} independent, standardized; diagonal variances often different from off-diagonal variances

Asymptotic Setting

- *Connecting dimension with sample size*
  - Suppose $\mathbf{X}$ is a $p \times n$ matrix with real- or complex-valued entries and independent columns
  - Specify that $p = p(n)$ and that

$$\lim_{n \to \infty} \frac{p}{n} = \gamma \in (0, \infty) \tag{3}$$

- *Wigner matrices*
  - Used as model for spectra of heavy atoms
  - Here $p = n$ such that $X_{ij} = \overline{X}_{ji}$ (symmetric/Hermitian; diagonal always real-valued)
  - X_{ij} independent, standardized; diagonal variances often different from off-diagonal variances
- *Wishart matrices*
  - Naturally arise as $\mathbf{X}\mathbf{X}^\top$
  - Note again the close connection to $\mathbf{S} = n^{-1}\mathbf{X}\mathbf{X}^\top$

How to Study Eigenvalues

- *Goal is to understand large-sample behavior of eigenvalues*
  - Eigenvalues of Wigner and Wishart matrices are real
  - But underlying matrix space is changing with p and n
  - No accumulation of degrees of freedom

How to Study Eigenvalues

- *Goal is to understand large-sample behavior of eigenvalues*
  - Eigenvalues of Wigner and Wishart matrices are real
  - But underlying matrix space is changing with p and n
  - No accumulation of degrees of freedom

- *Empirical spectral distribution (ESD)*
  - For any $N \times N$ matrix $\mathbf{Y}$ with eigenvalues $\lambda_1, \ldots, \lambda_N \in \mathbb{C}$ defined as $N^{-1} \sum_{\ell=1}^{N} \delta_{\lambda_\ell}$
  - For Hermitian $\mathbf{Y}$ this gives a mapping

$$F_{\mathbf{Y}} \colon \mathbb{R} \to [0,1], \; x \mapsto \frac{1}{N} \sum_{\ell=1}^{N} \mathbf{1}_{\{\lambda_\ell \leq x\}},$$

    called the ESD of $\mathbf{Y}$
  - The ESD is the fundamental object to conduct large-sample analysis in RMT
  - *Linear spectral statistics (LSS)* $\int g(x) dF_{\mathbf{Y}}(x)$ can be understood in terms of ESD

Spectrum of Sample Covariance Matrix

- *A simple example*
  - Take $n = 10$ observations of $p = 10$ dimensional centered Gaussian random vectors with identity population covariance matrix $\mathbf{\Sigma} = \mathbf{I}_{10}$
  - Population eigenvalues are $\ell_1 = \cdots = \ell_{10} = 1$
  - Sample eigenvalues $\hat{\ell}_1, \ldots, \hat{\ell}_{10}$ of $\mathbf{S}$ show an extreme spread
  - A typical sample would give

    $$0.003, \quad 0.036, \quad 0.095, \quad 0.160, \quad 0.300, \quad 0.510, \quad 0.780, \quad 1.120, \quad 1.400, \quad 3.070$$

    with variation over three orders of magnitude
  - Could conclude from sample that population eigenvalues are different from each other

Spectrum of Sample Covariance Matrix

- *A simple example*
  - Take $n = 10$ observations of $p = 10$ dimensional centered Gaussian random vectors with identity population covariance matrix $\mathbf{\Sigma} = \mathbf{I}_{10}$
  - Population eigenvalues are $\ell_1 = \cdots = \ell_{10} = 1$
  - Sample eigenvalues $\hat{\ell}_1, \ldots, \hat{\ell}_{10}$ of $\mathbf{S}$ show an extreme spread
  - A typical sample would give

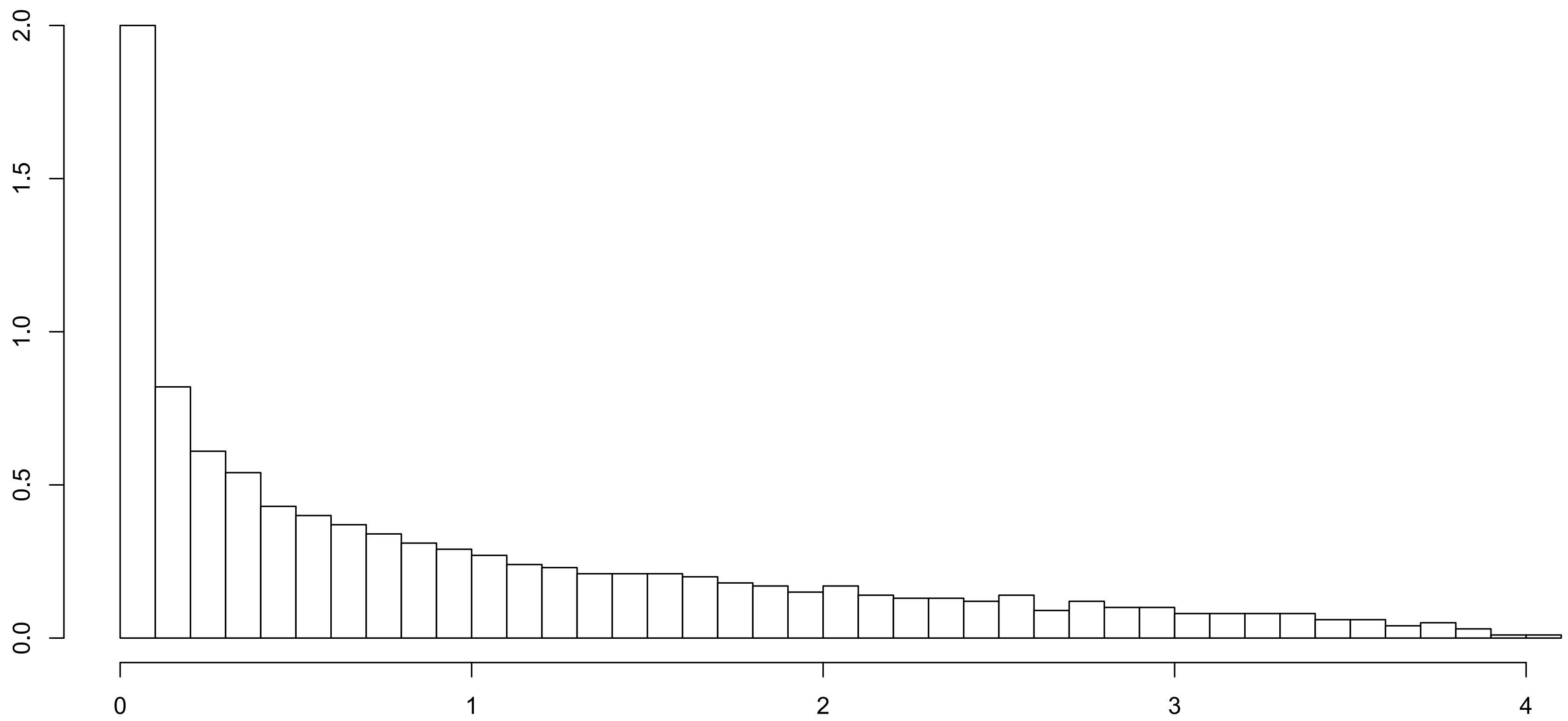
    $$0.003, \quad 0.036, \quad 0.095, \quad 0.160, \quad 0.300, \quad 0.510, \quad 0.780, \quad 1.120, \quad 1.400, \quad 3.070$$

    with variation over three orders of magnitude
  - Could conclude from sample that population eigenvalues are different from each other

- *Two immediate questions*
  - Does this phenomenon go away with larger n, p?
  - If not, what explains this disconnect between population and sample eigenvalues?

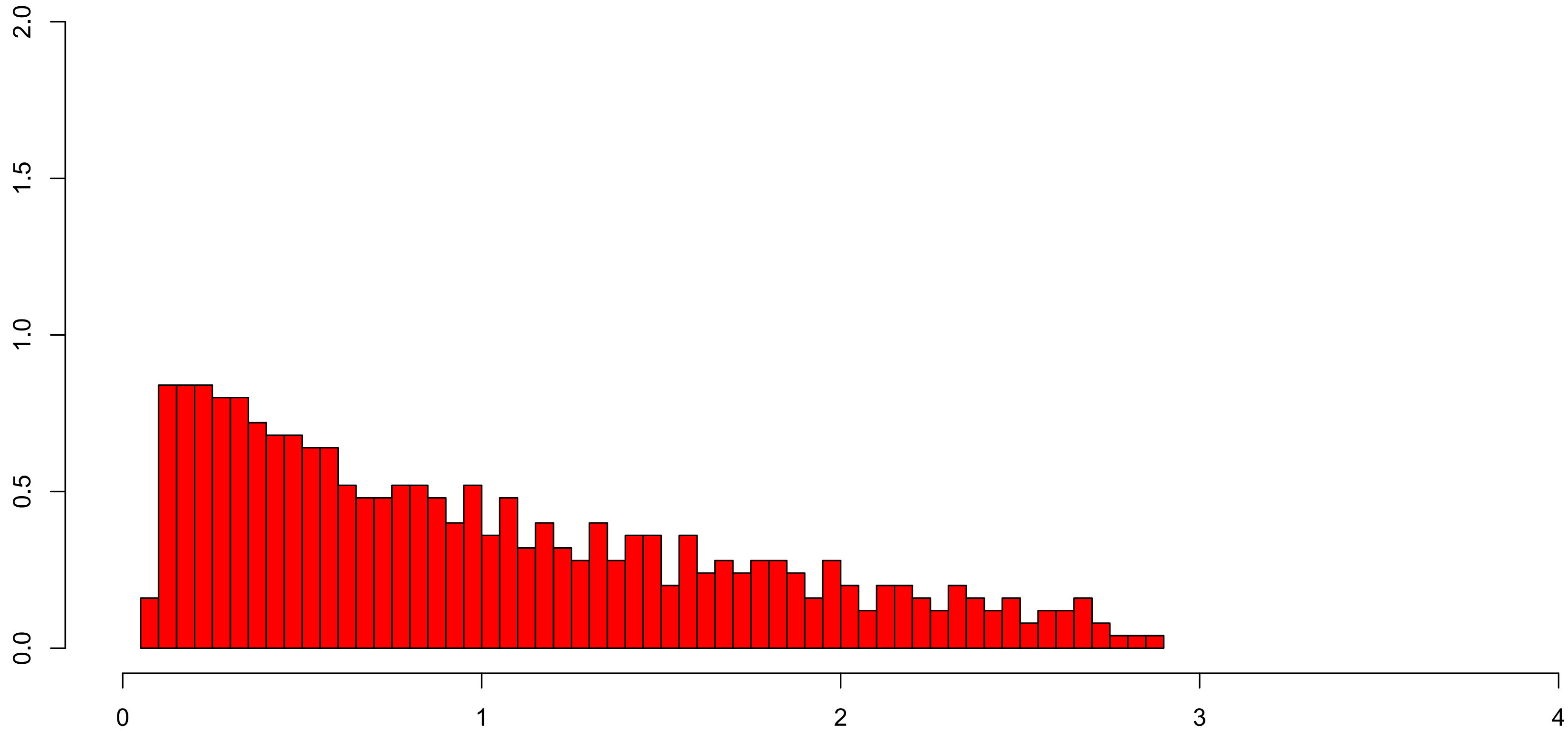
Spectrum of Sample Covariance Matrix

- Empirical spectrum of $\mathbf{S}$ for $n = 1000$ and $p = 1000$



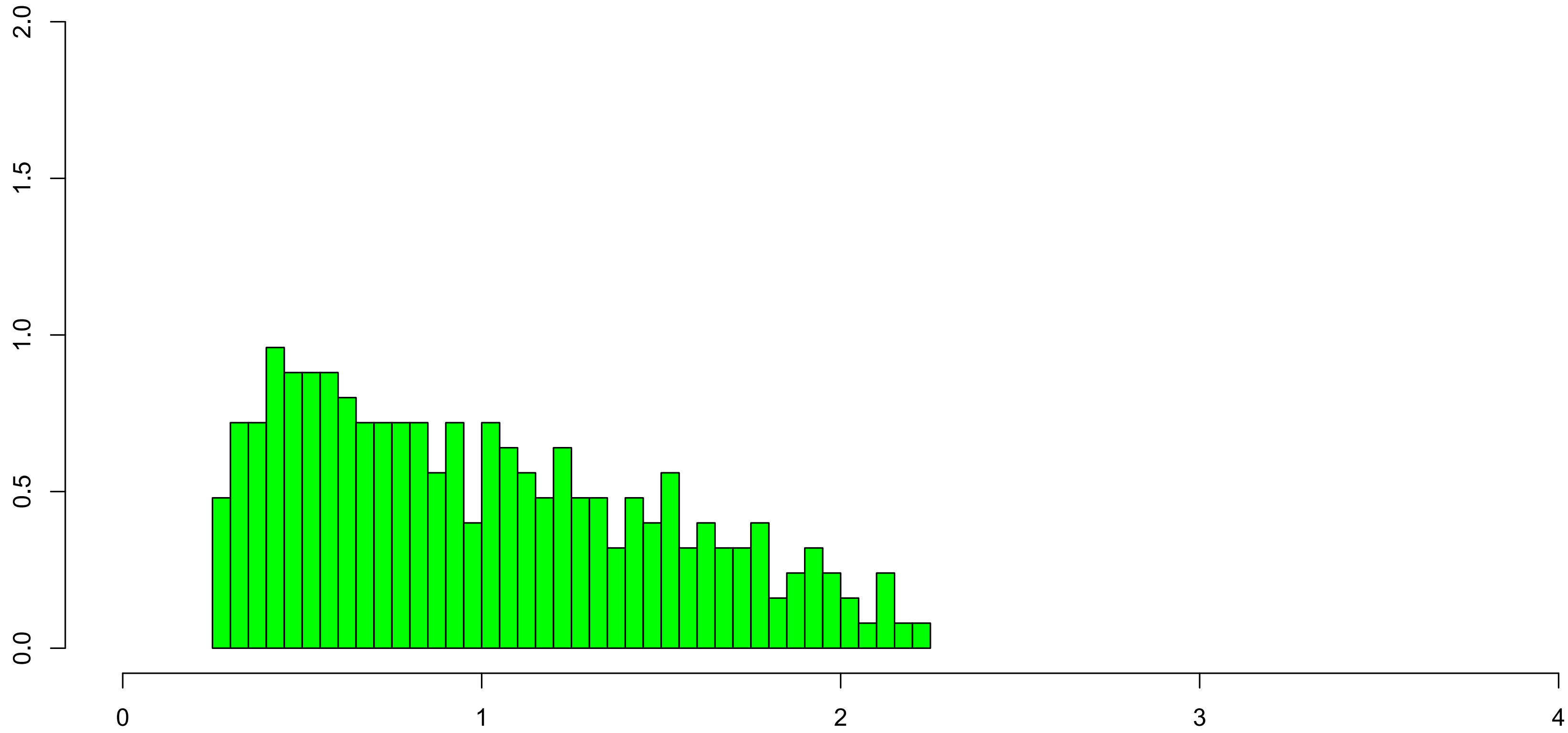
Spectrum of Sample Covariance Matrix

- Empirical spectrum of $\mathbf{S}$ for $n = 1000$ and $p = 500$



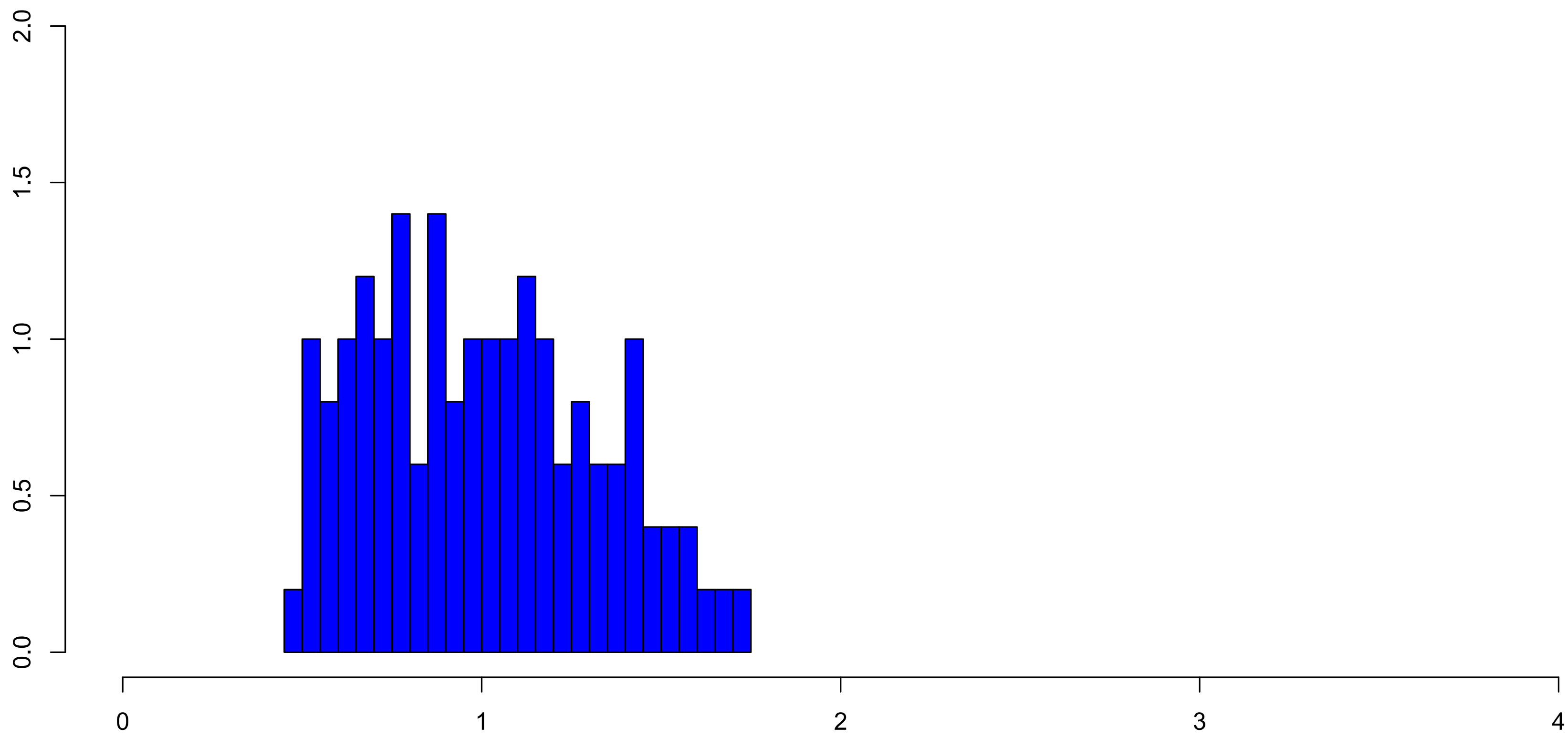
Spectrum of Sample Covariance Matrix

- Empirical spectrum of $\mathbf{S}$ for $n = 1000$ and $p = 250$



Spectrum of Sample Covariance Matrix

- Empirical spectrum of $\mathbf{S}$ for $n = 1000$ and $p = 100$



The Marčenko–Pastur Law

- *Assumptions*
  - Let $X_t = (X_{1t}, \ldots, X_{pt})^\top$, $t = 1, \ldots, n$, be observed
  - The entries X_{jt} are iid such that $\mathbb{E}[X_{11}] = 0$, $\mathbb{E}[|X_{11}|^2] = 1$ and $\mathbb{E}[|X_{11}|^4] < \infty$

The Marčenko–Pastur Law

- *Assumptions*
  - Let $X_t = (X_{1t}, \ldots, X_{pt})^\top$, $t = 1, \ldots, n$, be observed
  - The entries X_{jt} are iid such that $\mathbb{E}[X_{11}] = 0$, $\mathbb{E}[|X_{11}|^2] = 1$ and $\mathbb{E}[|X_{11}|^4] < \infty$
- *Under (3), the ESD $\hat{F}$ converges almost surely to a nonrandom limiting distribution F_γ*
  - If $\gamma \le 1$, the limiting distribution is continuous with density

$$f_\gamma(\lambda) = \frac{1}{2\pi\gamma}\sqrt{\frac{(b-\lambda)(\lambda-a)}{\lambda^2}}\mathbf{1}_{[a,b]}(\lambda),$$

    where $a = (1-\sqrt{\gamma})^2$ and $b = (1+\sqrt{\gamma})^2$
  - If $\gamma > 1$, the limiting distribution is a mixture of a point mass at 0 with weight $1 - 1/\gamma$ and the density f_γ with weight $1/\gamma$

The Marčenko–Pastur Law

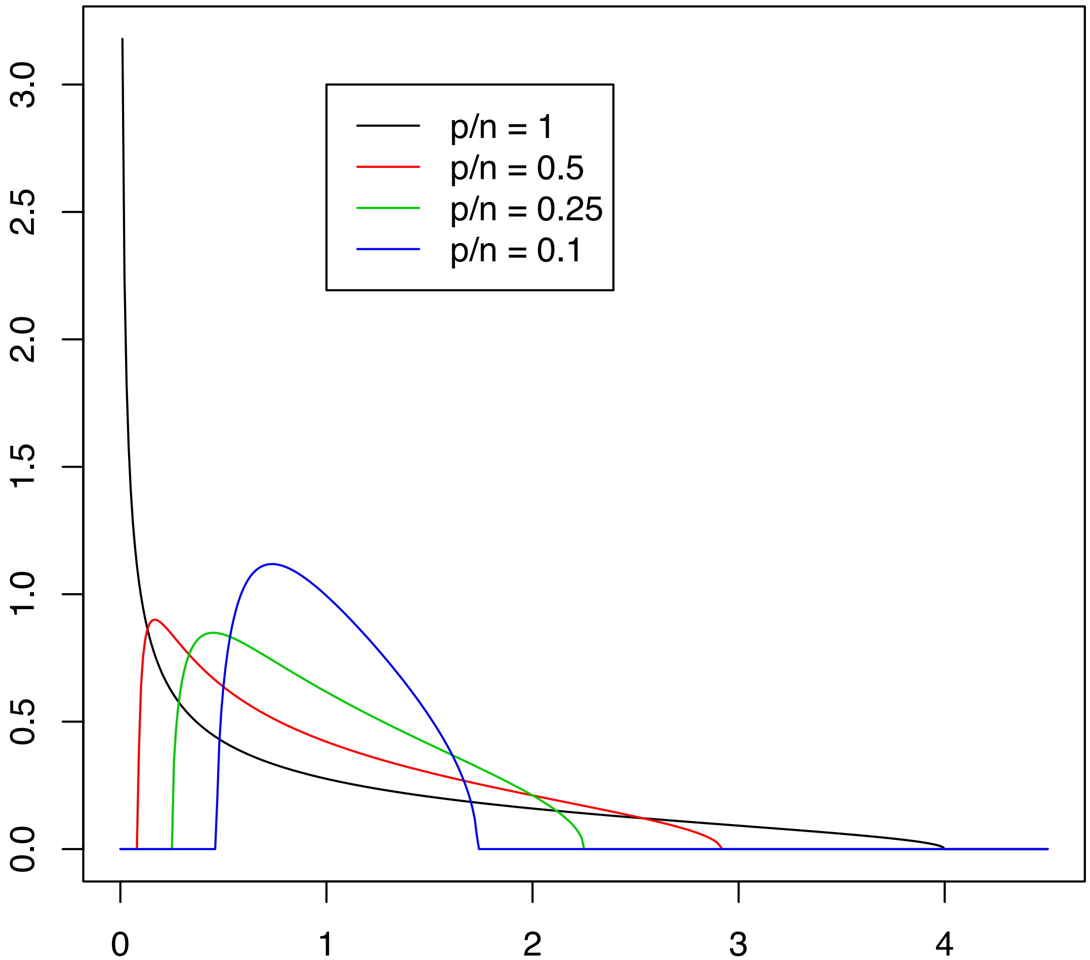
- *Assumptions*
  - Let $X_t = (X_{1t}, \ldots, X_{pt})^\top$, $t = 1, \ldots, n$, be observed
  - The entries X_{jt} are iid such that $\mathbb{E}[X_{11}] = 0$, $\mathbb{E}[|X_{11}|^2] = 1$ and $\mathbb{E}[|X_{11}|^4] < \infty$
- *Under (3), the ESD $\hat{F}$ converges almost surely to a nonrandom limiting distribution F_γ*
  - If $\gamma \le 1$, the limiting distribution is continuous with density

$$f_\gamma(\lambda) = \frac{1}{2\pi\gamma}\sqrt{\frac{(b-\lambda)(\lambda-a)}{\lambda^2}}\mathbf{1}_{[a,b]}(\lambda),$$

    where $a = (1-\sqrt{\gamma})^2$ and $b = (1+\sqrt{\gamma})^2$
  - If $\gamma > 1$, the limiting distribution is a mixture of a point mass at 0 with weight $1 - 1/\gamma$ and the density f_γ with weight $1/\gamma$
- *Consequences*
  - Spreading of the eigenvalues of $\mathbf{S}$ around the eigenvalues of $\mathbf{\Sigma}$ even in the limit
  - If $p/n \to 0$, the largest and smallest eigenvalue converge to 1 and classical results are retained

The Marčenko–Pastur Law

- MP law densities for different choices of $\gamma = \lim_{n\to\infty} \frac{p}{n}$



Stieltjes Transforms

- *Background*
  - Used extensively for determining limit behavior of ESD
  - Role in RMT similar to that of Fourier transform in probability theory

Stieltjes Transforms

- *Background*
  - Used extensively for determining limit behavior of ESD
  - Role in RMT similar to that of Fourier transform in probability theory

- *Definition and inversion formula*
  - The Stieltjes transform of measure μ on $\mathbb{R}$ is

$$s\colon \mathbb{C}_+ \to \mathbb{C}_+, \; z \mapsto \int \frac{1}{x-z} d\mu(x),$$

    where $\mathbb{C}^+ = \{z \in \mathbb{C} \colon \Im(z) > 0\}$ is the complex upper half-plane
  - s is analytic on $\mathbb{C}^+$
  - If $a < b$ are continuity points of a real probability measure μ, then

$$\mu(a,b] = \frac{1}{\pi} \lim_{v \to 0^+} \int_a^b \Im(s(u + \imath v)) du, \qquad z = u + \imath v$$

Resolvents and Stieltjes Transforms

- *Need the concept of resolvent*
  - Connection between sample covariance matrix $\mathbf{S}$, ESD $\hat{F}$ and Stieltjes transform $\hat{s} = s^{\hat{F}}$
  - The resolvent of $\mathbf{S}$ is

$$\mathbf{R}(z) = (\mathbf{S} - z\mathbf{I})^{-1}, \qquad z \in \mathbb{C}^+$$

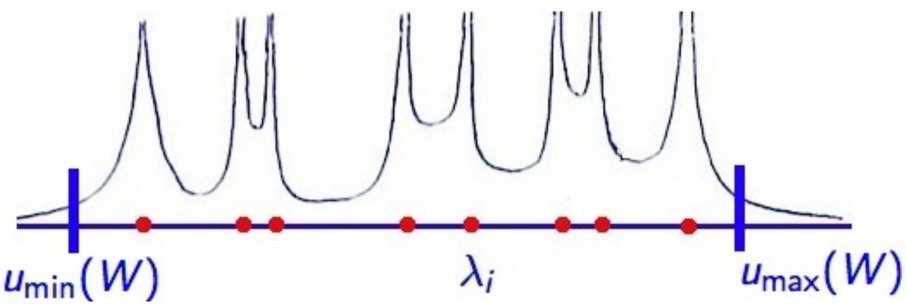
Resolvents and Stieltjes Transforms

- *Need the concept of resolvent*
  - Connection between sample covariance matrix $\mathbf{S}$, ESD $\hat{F}$ and Stieltjes transform $\hat{s} = s^{\hat{F}}$
  - The resolvent of $\mathbf{S}$ is

$$\mathbf{R}(z) = (\mathbf{S} - z\mathbf{I})^{-1}, \qquad z \in \mathbb{C}^+$$

- *Convergence of ESD through convergence of Stieltjes transform*
  - The Stieltjes transform of the ESD can be expressed as

$$\hat{s}(z) = \int \frac{1}{\lambda - z} d\hat{F}(\lambda) = \frac{1}{p}\sum_{j=1}^{p} \frac{1}{\lambda_j - z} = \frac{1}{p}\mathrm{tr}[(\mathbf{S} - z\mathbf{I})^{-1}] = \frac{1}{p}\mathrm{tr}[\mathbf{R}(z)]$$



High-Dimensional Time Series

- *Univariate and multivariate linear time series have been studied extensively*
  - Rather complete picture of strength and weaknesses of ARMA models
  - Many extensions available
  - Ready-to-use computer packages

High-Dimensional Time Series

- *Univariate and multivariate linear time series have been studied extensively*
  - Rather complete picture of strength and weaknesses of ARMA models
  - Many extensions available
  - Ready-to-use computer packages
- *Curse of dimensionality*
  - If dimension grows proportionally with the sample size, estimation becomes impossible
  - Assumptions on coefficient matrices are needed
  - Typically through imposing sparsity

High-Dimensional Time Series

- *Univariate and multivariate linear time series have been studied extensively*
  - Rather complete picture of strength and weaknesses of ARMA models
  - Many extensions available
  - Ready-to-use computer packages
- *Curse of dimensionality*
  - If dimension grows proportionally with the sample size, estimation becomes impossible
  - Assumptions on coefficient matrices are needed
  - Typically through imposing sparsity
- *Few results in the literature*
  - Review is given below
  - Existing contributions only touch the surface
  - Most of them are related to spectrum of sample covariance matrix

High-Dimensional Time Series

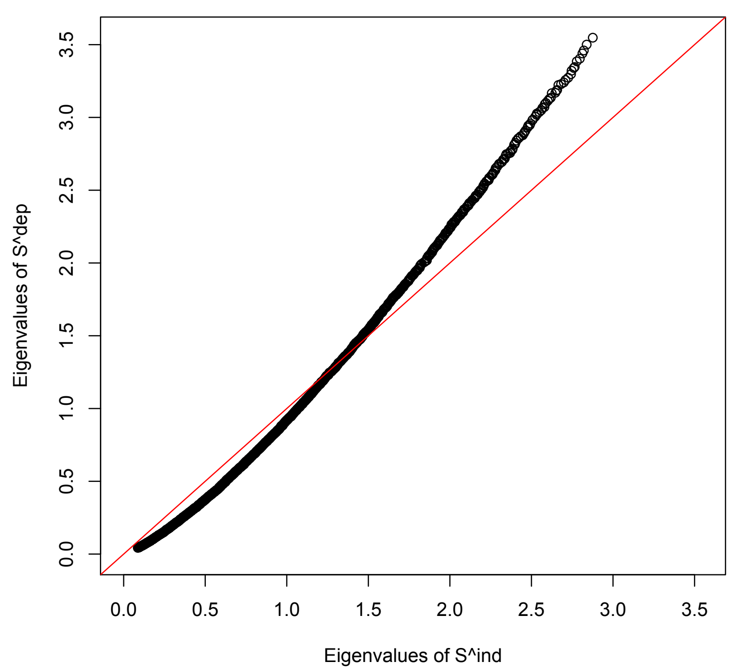
- Let Z_{jt} be standard normal and define the two processes $X_t^{\text{ind}} = Z_t$ and $X_t^{\text{dep}} = (Z_t + Z_{t-1})/\sqrt{2}$ as well as the sample covariance matrices $\mathbf{S}_{\text{ind}} = \frac{1}{n}\mathbf{X}^{\text{ind}}(\mathbf{X}^{\text{ind}})^*$ and $\mathbf{S}_{\text{dep}} = \frac{1}{n}\mathbf{X}^{\text{dep}}(\mathbf{X}^{\text{dep}})^*$

High-Dimensional Time Series

- Let Z_{jt} be standard normal and define the two processes $X_t^{\text{ind}} = Z_t$ and $X_t^{\text{dep}} = (Z_t + Z_{t-1})/\sqrt{2}$ as well as the sample covariance matrices $\mathbf{S}_{\text{ind}} = \frac{1}{n}\mathbf{X}^{\text{ind}}(\mathbf{X}^{\text{ind}})^*$ and $\mathbf{S}_{\text{dep}} = \frac{1}{n}\mathbf{X}^{\text{dep}}(\mathbf{X}^{\text{dep}})^*$

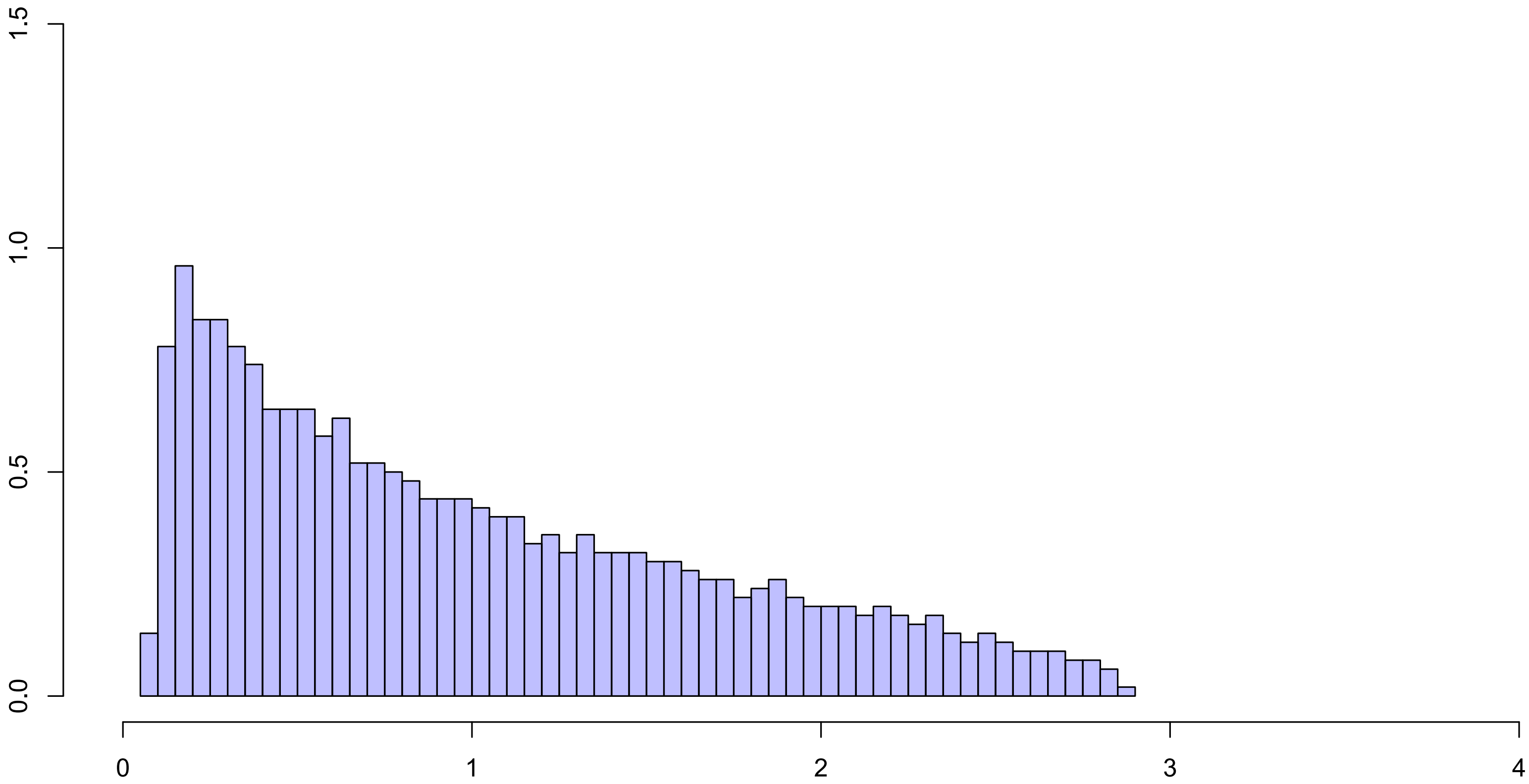
- Even though $\mathbb{E}[\mathbf{S}^{\text{ind}}] = \mathbb{E}[\mathbf{S}^{\text{dep}}]$, a comparison of eigenvalues (shown with $p = 1000$, $n = 2000$) reveals that the limiting behavior of the ESDs $\hat{F}^{\text{ind}}$ and $\hat{F}^{\text{dep}}$ is different

- How can this time series effect be quantified?



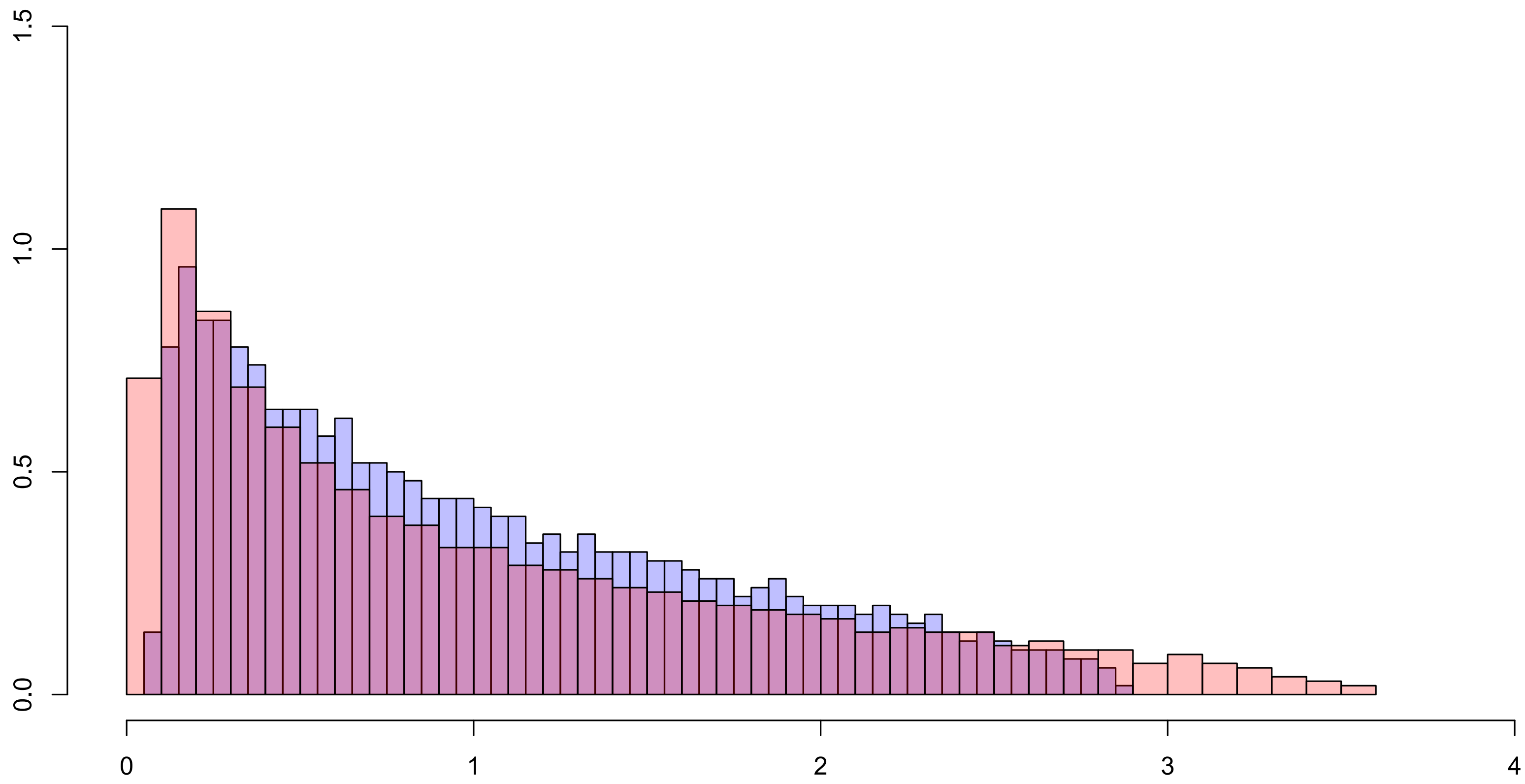
High-Dimensional Time Series

- Empirical spectrum of $\mathbf{S}$ for $n = 2000$ and $p = 1000$: Independent case



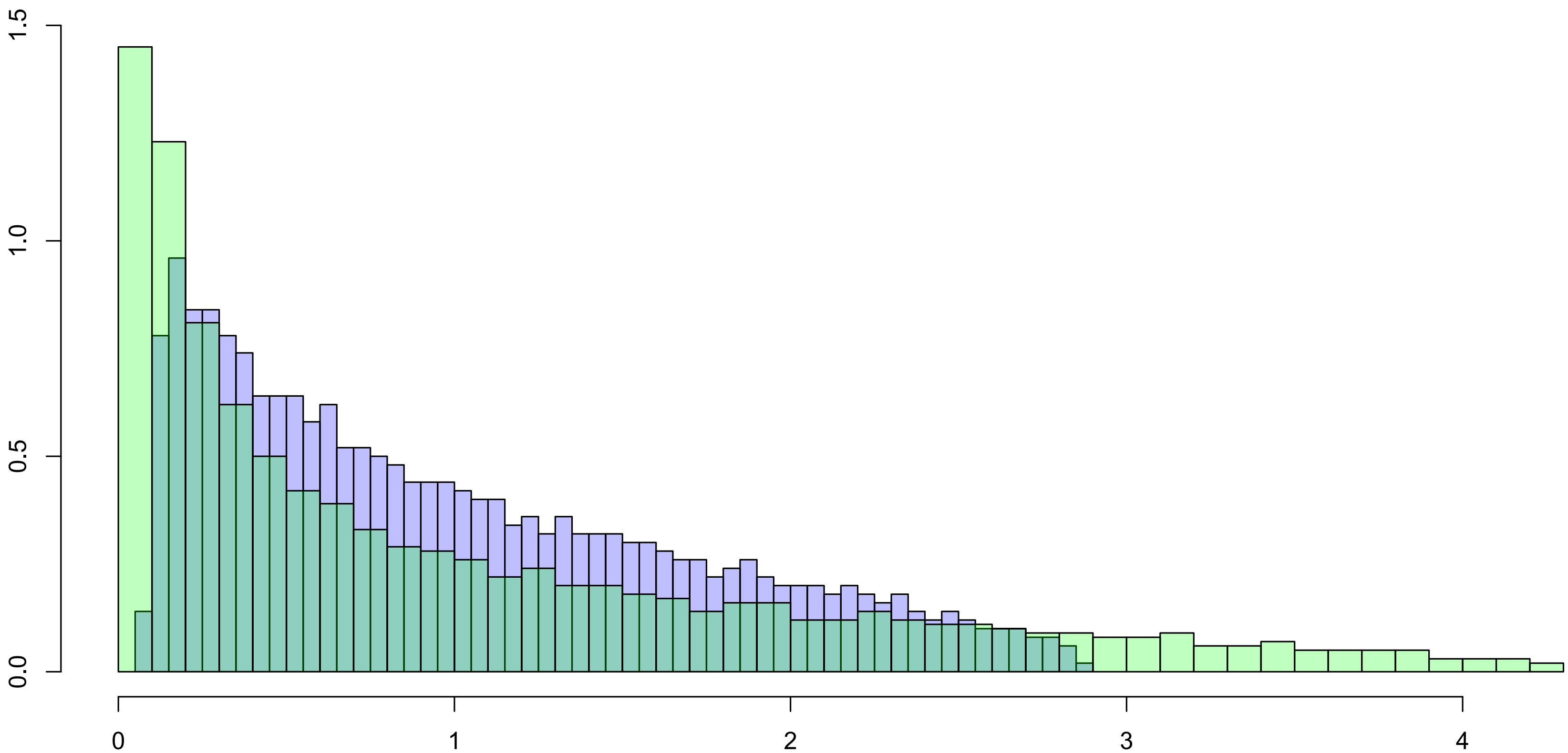
High-Dimensional Time Series

- Empirical spectrum of $\mathbf{S}$ for $n = 2000$ and $p = 1000$: Independent versus MA(1) case



High-Dimensional Time Series

- Empirical spectrum of $\mathbf{S}$ for $n = 2000$ and $p = 1000$: Independent versus MA(2) case



B. Spectral Theory for Linear Time Series

Outline

- Goal is to introduce framework that allows for

  — *description of linear processes in high-dimension*

  — *characterization of eigenvalues of sample covariance matrix*

  — *characterization of eigenvalues of symmetrized autocovariance matrices*

Literature Review

- Pfaffel and Schlemm (2011), *Theory of Probability and Mathematical Statistics* **31**, 313–329
  Yao (2012), *Statistics & Probability Letters* **82**, 22–28
  - Studied the linear time series model
    $$X_{jt} = \sum_{t'=0}^{\infty} \alpha_{t'} Z_{j,t-t'},$$
    with $(Z_{jt} \colon t \in \mathbb{Z}) \sim \mathrm{WN}(0,1)$ and independent rows $Z_1, \dots, Z_p$

Literature Review

- Pfaffel and Schlemm (2011), *Theory of Probability and Mathematical Statistics* **31**, 313–329
  Yao (2012), *Statistics & Probability Letters* **82**, 22–28
  - Studied the linear time series model
    $$X_{jt} = \sum_{t'=0}^{\infty} \alpha_{t'} Z_{j,t-t'},$$
    with $(Z_{jt} \colon t \in \mathbb{Z}) \sim \mathrm{WN}(0,1)$ and independent rows $Z_1, \ldots, Z_p$

- Jin et al. (2014), *The Annals of Applied Probability* **24**, 1199–1225
  - Studied the behavior of symmetrized autocovariance matrices in the independent case

Literature Review

- Pfaffel and Schlemm (2011), *Theory of Probability and Mathematical Statistics* **31**, 313–329
  Yao (2012), *Statistics & Probability Letters* **82**, 22–28
  - Studied the linear time series model

$$X_{jt} = \sum_{t'=0}^{\infty} \alpha_{t'} Z_{j,t-t'},$$

with $(Z_{jt} \colon t \in \mathbb{Z}) \sim \mathrm{WN}(0,1)$ and independent rows $Z_1, \ldots, Z_p$

- Jin et al. (2014), *The Annals of Applied Probability* **24**, 1199–1225
  - Studied the behavior of symmetrized autocovariance matrices in the independent case

- Hachem et al. (2005), *Markov Processes and Related Fields* **11**, 629–648
  - Studied the bi-stationary Gaussian process

$$X_{jt} = \sum_{j',t' \in \mathbb{Z}} h(j',t') Z_{j-j',t-t'},$$

with $h \in \ell^1(\mathbb{Z}^2)$ deterministic and $(Z_{jt} \colon j,t \in \mathbb{Z})$ iid real/complex standard normal

Assumptions for a Simple Time Series

- *Study first the MA(1) process* $X_t = Z_t + \mathbf{A}_1 Z_{t-1}$ satisfying

  (A1) $\mathbf{A}_1$ is a $p \times p$ Hermitian, possibly random, matrix independent of $(Z_t \colon t \in \mathbb{Z})$

  (A2) The ESD $F_p^{\mathbf{A}_1}$ of $\mathbf{A}_1$ converges weakly to a nonrandom probability distribution $F^{\mathbf{A}}$ (almost surely); there is $\bar{\lambda}_{\mathbf{A}} \geq 0$ such that $\|\mathbf{A}_1\| \leq \bar{\lambda}_{\mathbf{A}}$ (almost surely) for large p

Assumptions for a Simple Time Series

- *Study first the MA(1) process* $X_t = Z_t + \mathbf{A}_1 Z_{t-1}$ satisfying

  (A1) $\mathbf{A}_1$ is a $p \times p$ Hermitian, possibly random, matrix independent of $(Z_t \colon t \in \mathbb{Z})$

  (A2) The ESD $F_p^{\mathbf{A}_1}$ of $\mathbf{A}_1$ converges weakly to a nonrandom probability distribution $F^{\mathbf{A}}$ (almost surely); there is $\bar{\lambda}_{\mathbf{A}} \geq 0$ such that $\|\mathbf{A}_1\| \leq \bar{\lambda}_{\mathbf{A}}$ (almost surely) for large p

- *Motivation for assumptions*
  - Interest is in the spectrum of the covariance matrix $\mathbf{S}$
  - For an MA(1) process, we have $\mathbb{E}[\mathbf{S}] = \mathbf{I} + \mathbf{A}_1 \mathbf{A}_1^*$
  - The moments of the ESD of $\mathbf{S}$ depend on the trace of polynomials in $\mathbf{A}_1$, $\mathbf{A}_1^*$ and $\mathbf{A}_1 \mathbf{A}_1^*$
  - (A1) and (A2) ensure that the limiting ESD of $\mathbf{S}$ depends only on the limiting ESD of $\mathbf{A}_1$
  - Without these restrictions on $\mathbf{A}_1$, it is not clear what limit the ESD of $\mathbf{S}$ would have

Intuition for MA(1) Processes

- The limiting Stieltjes transform of $\hat{F}$ (ESD of $\mathbf{S}$) involves

$$h(\lambda, \nu) = 1 + 2\cos(\nu)\lambda + \lambda^2, \qquad \nu \in [0, 2\pi], \lambda \in \mathbb{R},$$

  - $h(\lambda, \cdot)$ is (up to normalization) the spectrum of the scalar MA(1) process $(x_t : t \in \mathbb{Z})$ given by $x_t = z_t + \lambda z_{t-1}$, $t \in \mathbb{Z}$

Intuition for MA(1) Processes

- The limiting Stieltjes transform of $\hat{F}$ (ESD of $\mathbf{S}$) involves

$$h(\lambda, \nu) = 1 + 2\cos(\nu)\lambda + \lambda^2, \qquad \nu \in [0, 2\pi], \lambda \in \mathbb{R},$$

  - $h(\lambda, \cdot)$ is (up to normalization) the spectrum of the scalar MA(1) process $(x_t \colon t \in \mathbb{Z})$ given by $x_t = z_t + \lambda z_{t-1}$, $t \in \mathbb{Z}$

- The limiting Stieltjes transform of the ESD $\hat{F}$ is determined from the Stieltjes kernel

$$K(z, \nu) = s^{(0)}(z) + 2\cos(\nu) s^{(1)}(z) + s^{(2)}(z), \qquad z \in \mathbb{C}^+, \nu \in [0, 2\pi],$$

  where

  - $s^{(k)}(z) = \lim_{n\to\infty} \frac{1}{p}\mathrm{tr}[(\mathbf{S} - z\mathbf{I})^{-1}\mathbf{A}_1^k]$, $k = 0, 1, 2$, where the limits exist in an a.s. sense
  - $s(z) = s^{(0)}(z)$ is the limiting Stieltjes transform of $\hat{F}$

Result for MA(1) Processes

Theorem 1: *Suppose the MA(1) process* $(X_t\colon t \in \mathbb{Z})$ *satisfies assumptions (A1) and (A2). Then, almost surely,* $\hat{F}$ *converges in distribution to a nonrandom probability distribution* F *with Stieltjes transform* $s(z)$ *given by*

$$s(z) = \int \left[\frac{1}{2\pi}\int_0^{2\pi} \frac{h(\lambda,\nu)}{1+cK(z,\nu)}d\nu - z\right]^{-1} dF^{\mathbf{A}}(\lambda), \tag{4}$$

where $K(z,\nu)$ *is the unique solution to the nonlinear equation*

$$K(z,\nu) = \int h(\lambda,\nu)\left[\frac{1}{2\pi}\int_0^{2\pi} \frac{h(\lambda,\nu')}{1+cK(z,\nu')}d\nu' - z\right]^{-1} dF^{\mathbf{A}}(\lambda), \tag{5}$$

for $\nu \in [0,2\pi]$*, with* $K(z,\nu)$ *satisfying the requirement that, for any* $\nu \in [0,2\pi]$*, it is the Stieltjes transform of a measure on* $\mathbb{R}$ *with total mass* $\int h(\lambda,\nu)dF^{\mathbf{A}}(\lambda)$.

Proof 1: Transformation to Independence

- Assume Gaussianity of $Z_1, \ldots, Z_n$

Proof 1: Transformation to Independence

- Assume Gaussianity of $Z_1, \ldots, Z_n$

- Let $\mathbf{L} = [o : e_1 : \cdots : e_{n-1}]$ and $\tilde{\mathbf{L}} = [e_n : e_1 : \cdots : e_{n-1}]$ be the $n \times n$ lag operator and its approximating circulant matrix, respectively, where o denotes the n-dimensional zero vector and e_j the jth canonical unit vector. Then, with $\mathbf{X} = [X_1 : \cdots : X_n]$ and $\mathbf{Z} = [Z_1 : \cdots : Z_n]$,

$$\mathbf{X} = \mathbf{Z} + \mathbf{A}_1\mathbf{Z}\mathbf{L} \qquad \text{and} \qquad \mathbf{X}_1 = \mathbf{Z} + \mathbf{A}_1\mathbf{Z}\tilde{\mathbf{L}},$$

  where $\mathbf{X}_1$ is a redefinition of $\mathbf{X}$ such that only the first column is changed to $Z_1 + \mathbf{A}_1 Z_n$

Proof 1: Transformation to Independence

- Assume Gaussianity of $Z_1, \ldots, Z_n$
- Let $\mathbf{L} = [o : e_1 : \cdots : e_{n-1}]$ and $\tilde{\mathbf{L}} = [e_n : e_1 : \cdots : e_{n-1}]$ be the $n \times n$ lag operator and its approximating circulant matrix, respectively, where o denotes the n-dimensional zero vector and e_j the jth canonical unit vector. Then, with $\mathbf{X} = [X_1 : \cdots : X_n]$ and $\mathbf{Z} = [Z_1 : \cdots : Z_n]$,

$$\mathbf{X} = \mathbf{Z} + \mathbf{A}_1\mathbf{Z}\mathbf{L} \qquad \text{and} \qquad \mathbf{X}_1 = \mathbf{Z} + \mathbf{A}_1\mathbf{Z}\tilde{\mathbf{L}},$$

  where $\mathbf{X}_1$ is a redefinition of $\mathbf{X}$ such that only the first column is changed to $Z_1 + \mathbf{A}_1 Z_n$
- Since $\tilde{\mathbf{L}}$ is a circulant matrix, it diagonalizes in the complex Fourier basis $\mathbf{U}_{\tilde{\mathbf{L}}}$.

Proof 1: Transformation to Independence

- Assume Gaussianity of $Z_1, \ldots, Z_n$

- Let $\mathbf{L} = [o : e_1 : \cdots : e_{n-1}]$ and $\tilde{\mathbf{L}} = [e_n : e_1 : \cdots : e_{n-1}]$ be the $n \times n$ lag operator and its approximating circulant matrix, respectively, where o denotes the n-dimensional zero vector and e_j the jth canonical unit vector. Then, with $\mathbf{X} = [X_1 : \cdots : X_n]$ and $\mathbf{Z} = [Z_1 : \cdots : Z_n]$,

$$\mathbf{X} = \mathbf{Z} + \mathbf{A}_1\mathbf{Z}\mathbf{L} \qquad \text{and} \qquad \mathbf{X}_1 = \mathbf{Z} + \mathbf{A}_1\mathbf{Z}\tilde{\mathbf{L}},$$

  where $\mathbf{X}_1$ is a redefinition of $\mathbf{X}$ such that only the first column is changed to $Z_1 + \mathbf{A}_1 Z_n$

- Since $\tilde{\mathbf{L}}$ is a circulant matrix, it diagonalizes in the complex Fourier basis $\mathbf{U}_{\tilde{\mathbf{L}}}$.

- Rotating with $\mathbf{U}_{\tilde{\mathbf{L}}}$ and using $\tilde{\mathbf{Z}} = [\tilde{Z}_1 : \cdots : \tilde{Z}_n] = \mathbf{Z}\mathbf{U}_{\tilde{\mathbf{L}}}$, the observations are transformed again into independent vectors $\tilde{X}_1, \ldots, \tilde{X}_n$ given by

$$\tilde{\mathbf{X}} = [\tilde{X}_1 : \cdots : \tilde{X}_n] = \mathbf{X}_1\mathbf{U}_{\tilde{\mathbf{L}}} = \big[(\mathbf{I} + \eta_1\mathbf{A}_1)\tilde{Z}_1 : \cdots : (\mathbf{I} + \eta_n\mathbf{A}_1)\tilde{Z}_n\big],$$

  where $\eta_t = e^{i\nu_t}$ and $\nu_t = 2\pi t/n$

Assumptions for Linear Processes

- Results for MA(q) processes can be proved as above, so focus on the MA(∞) process $(X_t\colon t \in \mathbb{Z})$ given by $X_t = \sum_{t'=0}^{\infty} \mathbf{A}_{t'} Z_{t-t'}$, let $\mathbf{A} = [\mathbf{A}_0 : \mathbf{A}_1 : \cdots]$. Assume that

(A3) The matrices $(\mathbf{A}_t\colon t \in \mathbb{N}_0)$ are simultaneously diagonalizable random Hermitian matrices, independent of $(Z_t\colon t \in \mathbb{Z})$ satisfying $\|\mathbf{A}_t\| \leq \bar{\lambda}_{\mathbf{A}_t}$ for all $t \in \mathbb{N}_0$ and large p with

$$\sum_{t=0}^{\infty} \bar{\lambda}_{\mathbf{A}_t} \leq \bar{\lambda}_{\mathbf{A}} < \infty \qquad \text{and} \qquad \sum_{t=0}^{\infty} t\bar{\lambda}_{\mathbf{A}_t} \leq \bar{\lambda}'_{\mathbf{A}} < \infty$$

(A4) There are continuous functions $f_t\colon \mathbb{R}^m \to \mathbb{R}$, $t \in \mathbb{N}_0$, such that for every p there is a set of points $\lambda_1, \dots, \lambda_p \in \mathbb{R}^m$, not necessarily distinct, and a unitary $p \times p$ matrix $\mathbf{U}$ such that

$$f_0(\lambda) = 1 \qquad \text{and} \qquad \mathbf{U}^*\mathbf{A}_t\mathbf{U} = \mathrm{diag}(f_t(\lambda_1), \dots, f_t(\lambda_p)), \qquad \ell \in \mathbb{N}$$

(A5) Almost surely, $F_p^{\mathbf{A}}$, the ESD of $\lambda_1, \dots, \lambda_p$, converges weakly to a nonrandom probability distribution function $F^{\mathbf{A}}$

Discussion of Assumptions

- *Simultaneous diagonizability can be relaxed to assuming Toeplitz structure for* $\mathbf{A}_\ell$ *with entries decaying away from the diagonal at an appropriate rate*

Discussion of Assumptions

- *Simultaneous diagonizability can be relaxed to assuming Toeplitz structure for $\mathbf{A}_\ell$ with entries decaying away from the diagonal at an appropriate rate*

- *ARMA(1,1) Example:* Let $(X_t \colon t \in \mathbb{Z})$ be given by

$$\mathbf{\Phi}(L)X_t = \mathbf{\Theta}(L)Z_t, \qquad t \in \mathbb{Z},$$

where $\mathbf{\Phi}(L) = \mathbf{I} - \mathbf{\Phi}_1 L$, $\mathbf{\Theta}(L) = \mathbf{I} + \mathbf{\Theta}_1 L$ such that $\|\mathbf{\Phi}_1\| \leq \bar{\phi} < 1$ and $\|\mathbf{\Theta}_1\| \leq \bar{\theta} < \infty$, and $(Z_t \colon t \in \mathbb{Z}) \sim \mathrm{IID}(0, \mathbf{I})$ with finite fourth moments. Then

  - $X_t = \mathbf{A}(L)Z_t$ with $\mathbf{A}(L) = \sum_{\ell=0}^\infty \mathbf{A}_\ell L^\ell = \mathbf{\Phi}^{-1}(L)\mathbf{\Theta}(L)$

  - Under simultaneous diagonizability, $\mathbf{U}\mathbf{\Phi}_1\mathbf{U}^* = \mathbf{\Lambda}_{\mathbf{\Phi}}$ and $\mathbf{U}\mathbf{\Theta}_1\mathbf{U}^* = \mathbf{\Lambda}_{\mathbf{\Theta}}$ with appropriate matrices $\mathbf{\Lambda}_{\mathbf{\Phi}} = \mathrm{diag}(\phi_1, \ldots, \phi_p)$ and $\mathbf{\Lambda}_{\mathbf{\Theta}} = \mathrm{diag}(\theta_1, \ldots, \theta_p)$ such that $|\phi_j| \leq \bar{\phi}$ and $|\theta_j| \leq \bar{\theta}$

  - Each coordinate of the rotated process satisfies

$$\frac{1 + \theta_j L}{1 - \phi_j L} = (1 + \theta_j L)\sum_{\ell=0}^\infty (\phi_j L)^\ell = 1 + (\theta_j + \phi_j)\sum_{\ell=1}^\infty \phi_j^{\ell-1} L^\ell,$$

  and it follows that $\mathbf{A}_\ell = \mathbf{U}\mathrm{diag}(f_\ell(\lambda_1), \ldots, f_\ell(\lambda_p))\mathbf{U}^*$ with $\lambda_j = (\phi_j, \theta_j)' \in \mathbb{R}^2$, $f_0(\lambda_j) = 1$ and $f_\ell(\lambda_j) = (\theta_j + \phi_j)\phi_j^{\ell-1}$ for $\ell \in \mathbb{N}$

Result for Linear Processes

- Define $\psi(\lambda,\nu) = \sum_{\ell=0}^{\infty} e^{i\ell\nu} f_\ell(\lambda)$ and $h(\lambda,\nu) = |\psi(\lambda,\nu)|^2$

Theorem 2: *If the linear process $(X_t \colon t \in \mathbb{Z})$ satisfies (A3)–(A5), then, almost surely, $\hat{F}$ converges weakly to a probability distribution F with Stieltjes transform $s(z)$ determined by the equation*

$$s(z) = \int \left[\frac{1}{2\pi} \int_0^{2\pi} \frac{h(\lambda,\nu)}{1 + cK(z,\nu)} d\nu - z \right]^{-1} dF^{\mathbf{A}}(\lambda), \tag{6}$$

where $K(z,\nu)$ is the unique solution to the nonlinear equation

$$K(z,\nu) = \int \left[\frac{1}{2\pi} \int_0^{2\pi} \frac{h(\lambda,\nu')}{1 + cK(z,\nu')} d\nu' - z \right]^{-1} h(\lambda,\nu)\, dF^{\mathbf{A}}(\lambda) \tag{7}$$

for $\nu \in [0, 2\pi]$, with $K(z,\nu)$ satisfying the requirement that, for any $\nu \in [0, 2\pi]$, it is the Stieltjes transform of a measure on $\mathbb{R}$ with total mass $\int h(\lambda,\nu) dF^{\mathbf{A}}(\lambda)$.

- Extensions to symmetrized autocovariance matrices exist

Examples

- If $\mathbf{A}_t = \mathbf{0}$, $t \in \mathbb{N}$, then $h(\lambda, \nu) \equiv 1$ and (6) reduces to the original Marčenko–Pastur law

Examples

- If $\mathbf{A}_t = \mathbf{0}$, $t \in \mathbb{N}$, then $h(\lambda, \nu) \equiv 1$ and (6) reduces to the original Marčenko–Pastur law

- If $\mathbf{A}_t = \alpha_t \mathbf{I}_p$, $t \in \mathbb{N}$, with $\sum_{t=1}^{\infty} t|\alpha_t| < \infty$, then

$$h(\lambda, \nu) \equiv h(\nu) = |\sum_{t=0}^{\infty} e^{\mathrm{i}t\nu} \alpha_t|^2$$

is independent of λ and (6) reduces to

$$s(z) = \left[\frac{1}{2\pi} \int_0^{2\pi} \frac{h(\nu) d\nu}{1 + cs(z)h(\nu)} - z \right]^{-1}$$

that is, the linear process case with independent, identically distributed rows

Pfaffel and Schlemm (2011), *Probability and Mathematical Statistics* **31**, 313–329

Yao (2012), *Statistics & Probability Letters* **82**, 22–28

Examples

- If $\mathbf{A}_t = \mathbf{0}$, $t \in \mathbb{N}$, then $h(\lambda, \nu) \equiv 1$ and (6) reduces to the original Marčenko–Pastur law

- If $\mathbf{A}_t = \alpha_t \mathbf{I}_p$, $t \in \mathbb{N}$, with $\sum_{t=1}^{\infty} t|\alpha_t| < \infty$, then

$$h(\lambda, \nu) \equiv h(\nu) = |\sum_{t=0}^{\infty} e^{\mathrm{i}t\nu} \alpha_t|^2$$

  is independent of λ and (6) reduces to

$$s(z) = \left[\frac{1}{2\pi} \int_0^{2\pi} \frac{h(\nu)d\nu}{1 + cs(z)h(\nu)} - z \right]^{-1}$$

  that is, the linear process case with independent, identically distributed rows

  Pfaffel and Schlemm (2011), *Probability and Mathematical Statistics* **31**, 313–329
  Yao (2012), *Statistics & Probability Letters* **82**, 22–28

- Causal ARMA processes included by determining the causal matrix coefficients

Final Comments on the Proof

- *Arguments used so far do not work because*
  - if one constructs the data matrix $\mathbf{X}$ not from a linear process $X_t = \sum_{t'=0}^{\infty} \mathbf{A}_{t'} Z_{t-t'}$, then every column of $\mathbf{X}$ is different from the transformed matrix $\mathbf{X}_\infty = \sum_{t'=0}^{\infty} \mathbf{A}_t \mathbf{Z} \tilde{\mathbf{L}}^t$ and not only the first column as in the MA(1) case
  - for the MA(1) case, one can write the Stieltjes transform $s_p(z)$ as a function of $2p(n+1)$ variables Z_{tj}^R and Z_{tj}^I, but for linear processes, even for finite p, $s_p(z)$ is a function of infinitely many Z_{tj}^R and Z_{tj}^I

Final Comments on the Proof

- *Arguments used so far do not work because*
  - if one constructs the data matrix $\mathbf{X}$ not from a linear process $X_t = \sum_{t'=0}^{\infty} \mathbf{A}_{t'} Z_{t-t'}$, then every column of $\mathbf{X}$ is different from the transformed matrix $\mathbf{X}_\infty = \sum_{t'=0}^{\infty} \mathbf{A}_t \mathbf{Z} \tilde{\mathbf{L}}^t$ and not only the first column as in the MA(1) case
  - for the MA(1) case, one can write the Stieltjes transform $s_p(z)$ as a function of $2p(n+1)$ variables Z_{tj}^R and Z_{tj}^I, but for linear processes, even for finite p, $s_p(z)$ is a function of infinitely many Z_{tj}^R and Z_{tj}^I

- *Use approximation through finite-order MA processes* $X_t^{q(p)} = \sum_{t'=0}^{q(p)} \mathbf{A}_{t'} Z_{t-t'}$ *whose order* $q(p)$ *is growing with the sample size*
  - Obviously $q(p) \to \infty$ is necessary
  - But $q(p)$ cannot grow too fast (same difficulties in transitioning from the Gaussian to the non-Gaussian case as for the linear process itself) or too slow (showing that the limiting ESDs of the linear process and its truncated version are the same becomes an issue)
  - Choose $q(p) = \lceil p^{1/4} \rceil$, with $\lceil \cdot \rceil$ denoting the ceiling function

C. Estimation of Quadratic Forms for Time Series

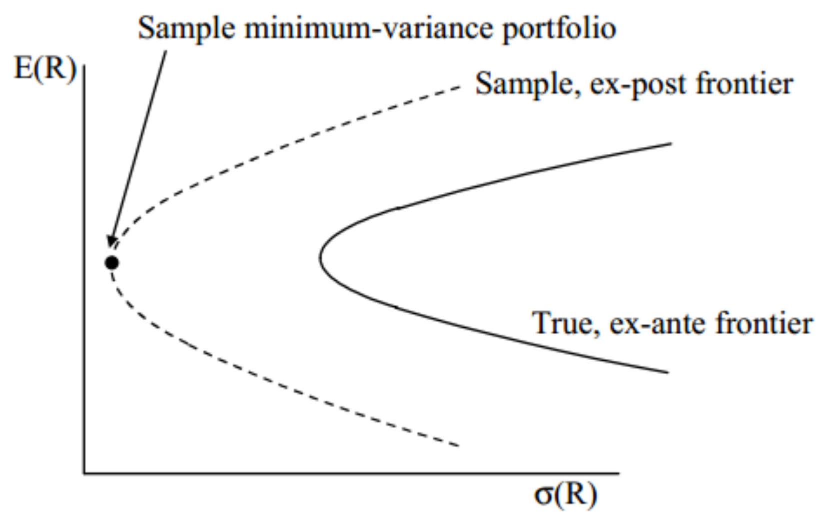
Outline

- Goal is to make framework more applicable

  — *Estimation of quadratic forms involving sample covariance matrices*

  — *Lead example: Markowitz portfolio and mean-variance frontier*

  — *Based on a thresholding and model selection procedure for eigenvalues*



Markowitz Portfolio Problem

- *Framework for assembling a portfolio of risky assets* $v_1, \ldots, v_p$
  - Assets have expected returns $\mu_1, \ldots \mu_p$ and covariance matrix $\mathbf{\Sigma}$
  - For expected portfolio return μ_P choose allocation with smallest risk

Markowitz Portfolio Problem

- *Framework for assembling a portfolio of risky assets* $v_1, \ldots, v_p$
  - Assets have expected returns $\mu_1, \ldots \mu_p$ and covariance matrix $\mathbf{\Sigma}$
  - For expected portfolio return μ_P choose allocation with smallest risk
- *Mathematical formulation as quadratic program*
  - Solve

$$\min_{w \in \mathbb{R}^p} \frac{1}{2} w' \mathbf{\Sigma} w$$

    with linear constraints $w'\mu = \mu_P$ and $w'\mathbf{1} = 1$, where $\mu = (\mu_1, \ldots, \mu_p)'$ and $\mathbf{1} = (1, \ldots, 1)'$
  - If w_{opt} is the solution, then $w'_{opt} \mathbf{\Sigma} w_{opt}$ viewed as function of μ_P is called *efficient frontier*
  - If Σ is invertible, then there is an explicit form of w_{opt}

Markowitz Portfolio Problem

- *Framework for assembling a portfolio of risky assets* $v_1, \ldots, v_p$
  - Assets have expected returns $\mu_1, \ldots \mu_p$ and covariance matrix $\mathbf{\Sigma}$
  - For expected portfolio return μ_P choose allocation with smallest risk

- *Mathematical formulation as quadratic program*
  - Solve
$$\min_{w \in \mathbb{R}^p} \tfrac{1}{2} w' \mathbf{\Sigma} w$$
    with linear constraints $w'\mu = \mu_P$ and $w'\mathbf{1} = 1$, where $\mu = (\mu_1, \ldots, \mu_p)'$ and $\mathbf{1} = (1, \ldots, 1)'$
  - If w_{opt} is the solution, then $w'_{opt}\mathbf{\Sigma} w_{opt}$ viewed as function of μ_P is called *efficient frontier*
  - If Σ is invertible, then there is an explicit form of w_{opt}

- *Common practice: Estimate the expected return vector* μ *and use* $\mathbf{S}$ *in place of* $\mathbf{\Sigma}$
  - This can lead to risk underestimation, especially when n and p are comparable
  - Results available in the high-dimensional setting are for independent setting

Risk Underestimation

!

- To highlight the differences between the optimal weights obtained from the population and sample quadratic programs, let

$$w = w_{opt,p} \qquad \text{and} \qquad \hat{w} = w_{opt,s}$$

the population and sample weights, respectively

Risk Underestimation

- To highlight the differences between the optimal weights obtained from the population and sample quadratic programs, let

$$w = w_{opt,p} \qquad \text{and} \qquad \hat{w} = w_{opt,s}$$

the population and sample weights, respectively

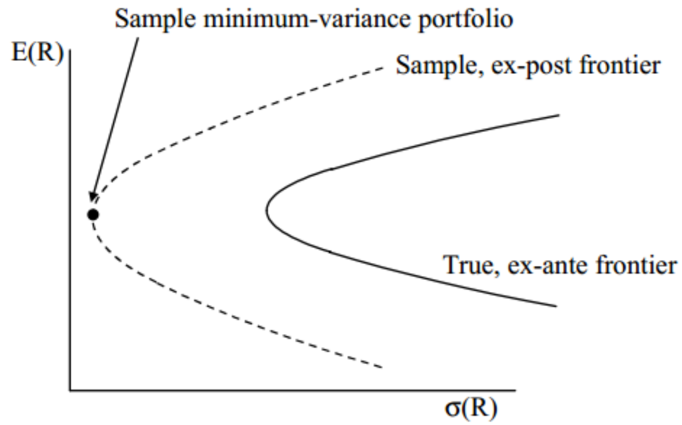
- Then, assuming $p \leq n$ for simplicity,

$$\hat{w}'\mathbf{S}^{-1}\hat{w} \approx N_p\left(w'\boldsymbol{\Sigma}^{-1}w - D_p\right) < w'\boldsymbol{\Sigma}^{-1}w,$$

where

$$N_p = 1 - \frac{p-2}{n-1},$$

$$D_d = \frac{p}{n}\left(u_P'\mathbf{Q}^{-1}e_2\right)^2\left(1 + \frac{p}{n}e_2'\mathbf{Q}^{-1}e_2\right)^{-1}$$



Algorithm: Idea

- *The eigendecomposition of* $\mathbf{\Sigma}$ *gives*

$$\mathbf{Q} = \mathbf{V}'\mathbf{\Sigma}^{-1}\mathbf{V} = \mathbf{V}'\mathbf{U}'\mathbf{\Lambda}^{-1}\mathbf{U}\mathbf{V},$$

Algorithm: Idea

- *The eigendecomposition of* $\mathbf{\Sigma}$ *gives*

$$\mathbf{Q} = \mathbf{V}'\mathbf{\Sigma}^{-1}\mathbf{V} = \mathbf{V}'\mathbf{U}'\mathbf{\Lambda}^{-1}\mathbf{U}\mathbf{V},$$

- *Perform the following steps:*

Step 1: To estimate $\mathbf{\Lambda}$, utilize that LSD is given by

$$s(z) = \int \left[\frac{1}{2\pi} \int_0^{2\pi} \frac{h(\lambda, \nu)}{1 + cK(z, \nu)} d\nu - z \right]^{-1} dF^{\mathbf{A}}(\lambda), \tag{8}$$

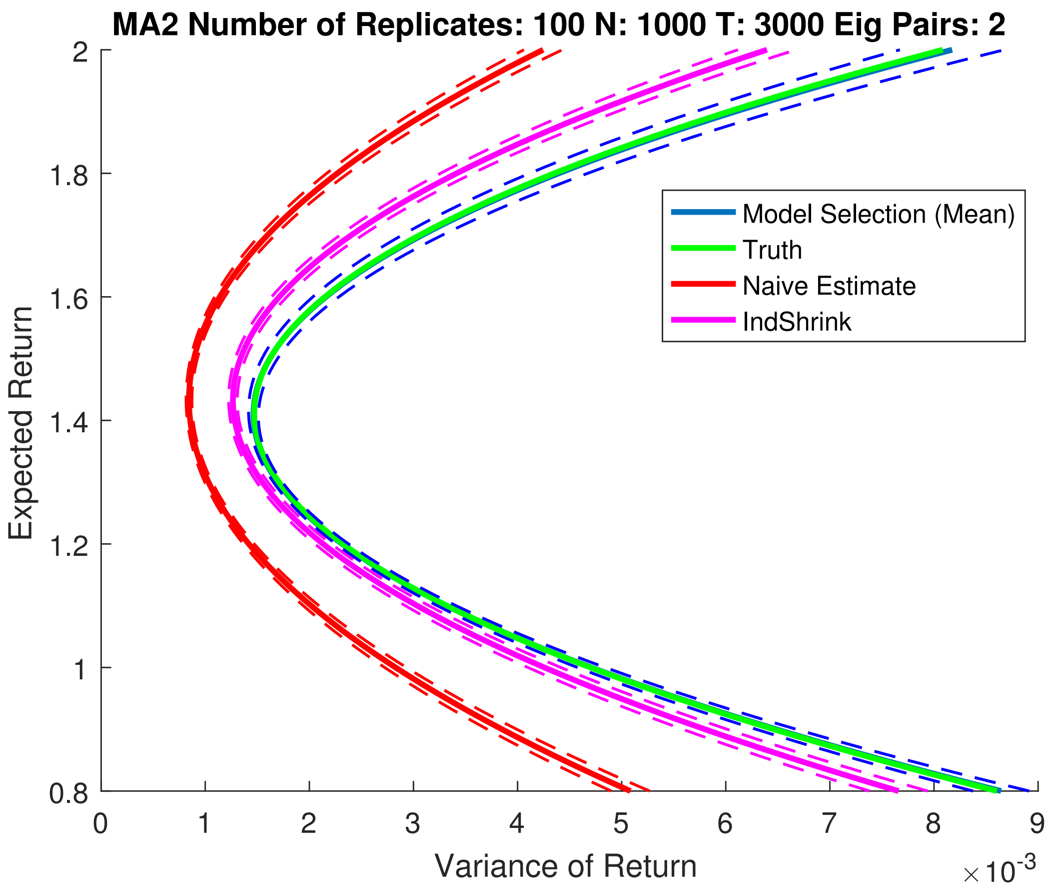
and mimic limiting behavior on sample version, using $\hat{s}(z)$ in place of $s(z)$

Step 2: Invert (8) to find $\hat{F}^{\mathbf{A}}$: Choose best-fitting spectrum from set of candidate spectra

Step 3: Estimate contribution of columns of $\mathbf{UV}$ using projection matrices

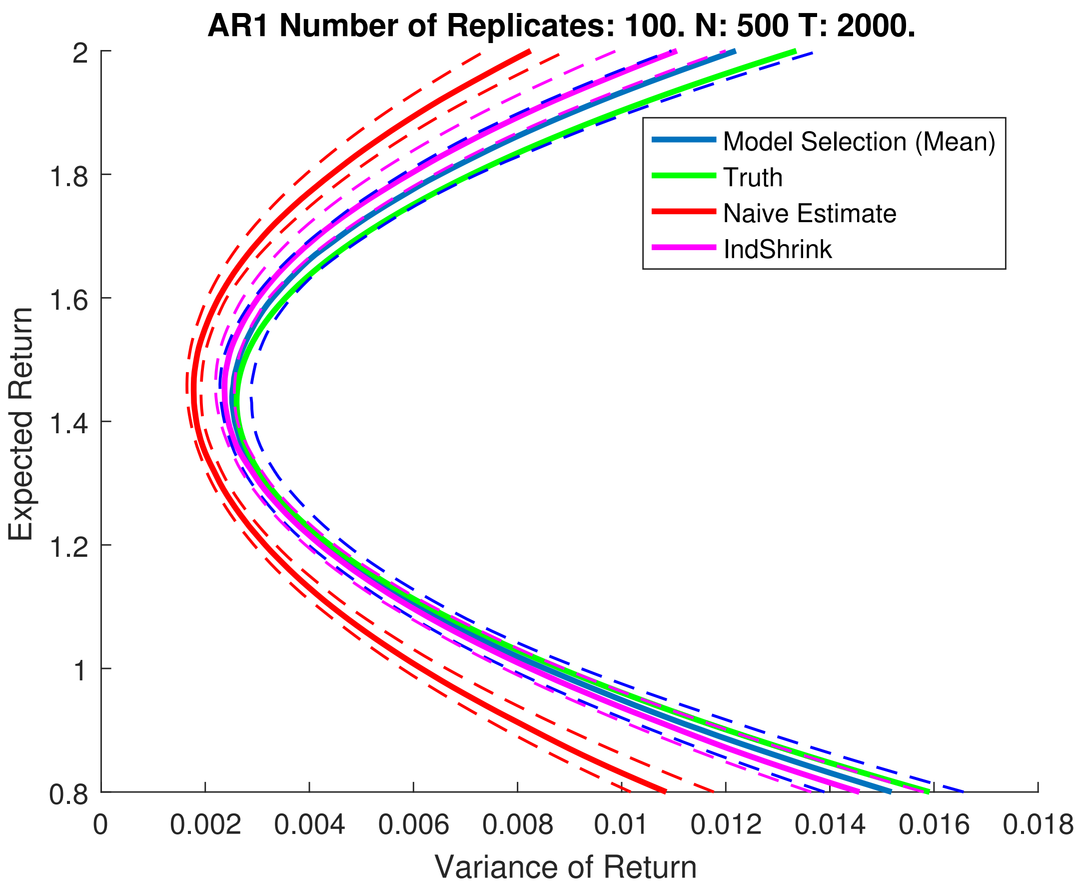
Performance: MA(2) Process

- $p = 1000$, $n = 3000$. *"Model Selection"* is proposed algorithm; *"Naive Estimate"* uses $\mathbf{S}$ in place of $\boldsymbol{\Sigma}$; *"IndShrink"* is shrinkage estimation assuming independence



Performance: AR(1) Process

- $p = 500$, $n = 2000$. Labeling is as before.
- Model misspecification: An AR(1) time series is approximated by an MA(2) time series



D. Wrap-Up

Wrap-Up of Talk

- *Learnt about*
  - the bulk eigenvalues of sample (auto)covariances from linear processes
  - the difficulties in finding appropriate models for high-dimensional time series
  - Some potential applications
  - One actual application: Mean-variance frontier estimation

- *Learnt also that much more work is needed*