Some remarks on eigenvalues of large random covariance matrices

Noureddine El Karoui
Statistics, UC Berkeley

July 10th, 2006

Sample covariance matrix and its eigenvalues

- Data: $n \times p$ matrix X
- n (independent identically distributed) observations of a random vector $(X_i)_{i=1}^n$ in $\mathbb{R}^p$.
- Suppose X_i-s have covariance matrix Σ_p

Sample covariance matrix and its eigenvalues

- Data: $n \times p$ matrix X
- n (independent identically distributed) observations of a random vector $(X_i)_{i=1}^n$ in $\mathbb{R}^p$.
- Suppose X_i-s have covariance matrix Σ_p;
- Problems of interest:
  1. **Testing** hypotheses, like $\Sigma_p = \mathrm{Id}_p$
  2. **Estimating** Σ_p (e.g for PCA); eigenvalues: λ_i

     Standard estimator: $\widehat{\Sigma}_p = (X - \bar{X})'(X - \bar{X})/(n-1)$: eigenvalues l_i

What do we need from theory?

- For **Testing**: want to test $\Sigma_p = \mathrm{Id}_p$ using $l_1(\widehat{\Sigma}_p)$ as our statistic. Need:
  1. Behavior of l_1 when $\Sigma_p = \mathrm{Id}_p$ (design test)
  2. Behavior of l_1 when $\Sigma_p \neq \mathrm{Id}_p$ (power)
  3. Speed of convergence if using asymptotics

- For **Estimation**:
  1. Relation between l_i's and λ_i's
  2. Method to invert this relation:go from l_i to λ_i

Classical results from multivariate Statistics

Classical theory: p fixed, $n \to \infty$

Classical results from multivariate Statistics

Classical theory: p fixed, $n \to \infty$
Assuming

- X_i are $\mathcal{N}(0, \Sigma_p)$
- Eigenvalues of Σ_p all have multiplicity 1

$$\forall k \leq p \, , \sqrt{n} \, (l_k(X'X/n) - \lambda_k) \Rightarrow \mathcal{N}(0, 2\lambda_k^2)$$

Result due to T. Anderson, '63

Much more is known about estimation of covariance matrices: Stein, Haff, etc...

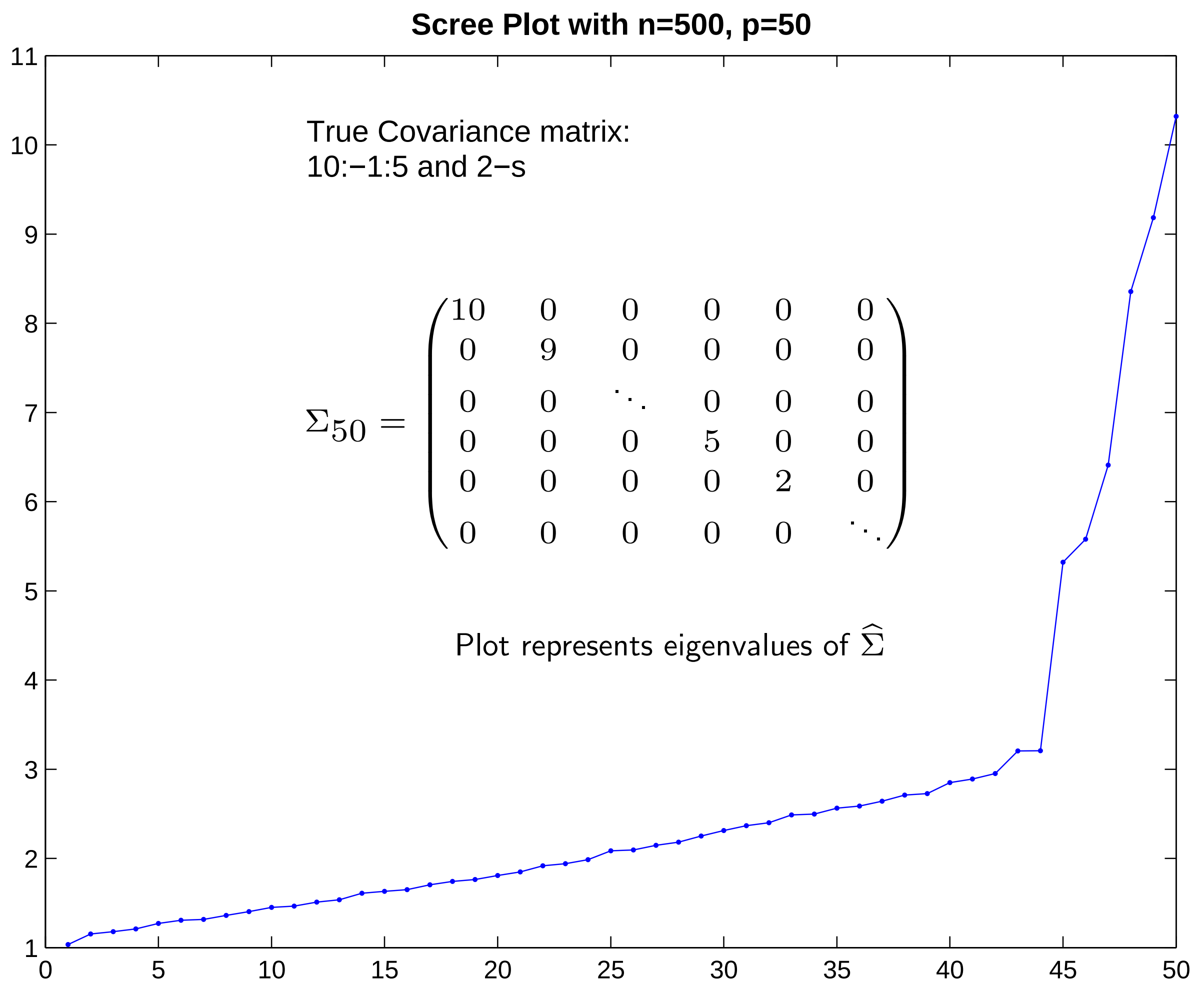
Nowadays: Large n, Large p

Modern datasets often exhibit large n and large p. Examples include data used for

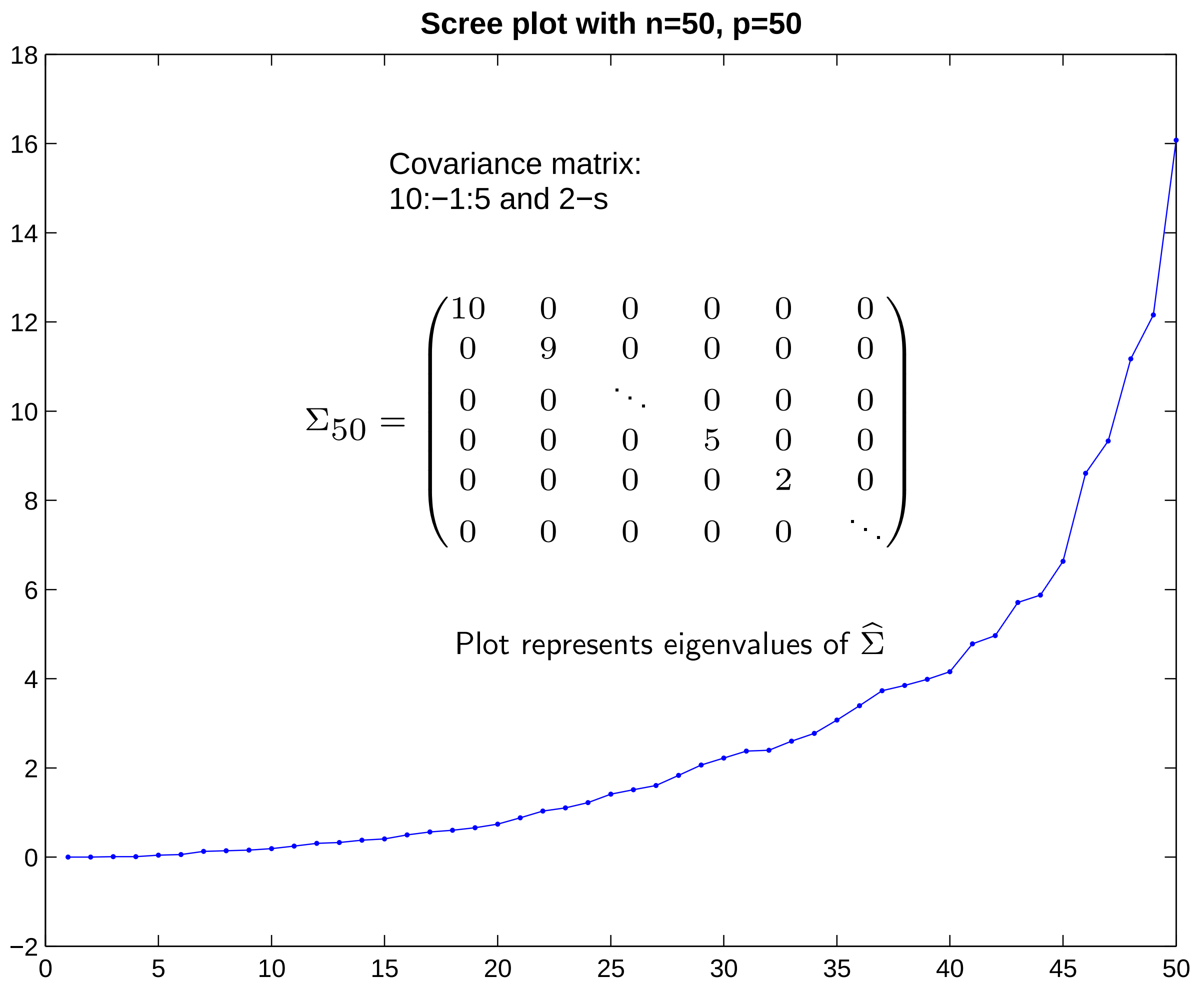
- **Speech Recognition** (n a few 100-s, p a few 100-s)
- **Face recognition**: Eigenfaces ($n = 400$, and $p \simeq 10,000$ for ORL face database)
- **Climate**: Complex PCA ($n = 320$, $p = 1440$)
- **Finance**: Industry portfolios observed for 2 years($n = 504$ and $p = 48$)

Cannot assume $n \gg p$ or p fixed. Ratio $\rho_n = p/n$ not always close to zero.

Scree plot with $n = 500$, $p = 50$



Importance of n/p: an example with $n = 50$, $p = 50$



Plan

1. **Estimation**
   - Classical result from Random Matrix: Marčenko-Pastur Equation
   - Numerical inversion of Marčenko-Pastur Equation: shrinkage of eigenvalues
   - A new estimator for eigenvalues

2. **Testing**
   - Role of Marčenko-Pastur equation in theory: fluctuation behavior of largest eigenvalue l_1 when $\Sigma_p \neq \mathrm{Id}_p$
   - Rates of convergence

Important adjustment

Change focus from high-dimensional vectors to associated measures

Important adjustment

Change focus from high-dimensional vectors to associated measures

$$(y_1, \ldots, y_p) \xrightarrow{\text{associated distribution}} dM_p(x) = \frac{1}{p} \sum_{i=1}^{p} \delta_{y_i}(x)$$

Important adjustment

Change focus from high-dimensional vectors to associated measures

$$(y_1, \ldots, y_p) \overset{\text{associated distribution}}{\longrightarrow} dM_p(x) = \frac{1}{p} \sum_{i=1}^{p} \delta_{y_i}(x)$$

For sample covariance matrix,

$$(l_1, \ldots, l_p) \longrightarrow \boldsymbol{dF_p(x)} = \frac{1}{p} \sum_{i=1}^{p} \delta_{l_i}(x) \ \ \textbf{empirical spectral distribution}$$

For population ("true") covariance matrix,

$$(\lambda_1, \ldots, \lambda_p) \longrightarrow \boldsymbol{dH_p(x)} = \frac{1}{p} \sum_{i=1}^{p} \delta_{\lambda_i}(x) \ \ \textbf{population spectral distribution}$$

Important adjustment

Change focus from high-dimensional vectors to associated measures

$$(y_1, \ldots, y_p) \overset{\text{associated distribution}}{\longrightarrow} dM_p(x) = \frac{1}{p}\sum_{i=1}^{p} \delta_{y_i}(x)$$

For sample covariance matrix,

$$(l_1, \ldots, l_p) \longrightarrow \boldsymbol{dF_p(x)} = \frac{1}{p}\sum_{i=1}^{p} \delta_{l_i}(x) \ \ \textbf{empirical spectral distribution}$$

For population ("true") covariance matrix,

$$(\lambda_1, \ldots, \lambda_p) \longrightarrow \boldsymbol{dH_p(x)} = \frac{1}{p}\sum_{i=1}^{p} \delta_{\lambda_i}(x) \ \ \textbf{population spectral distribution}$$

Estimation **AIM**: From F_p get $\widehat{H}_p$ that approximates well H_p

Interlude: Stieltjes transforms of distributions

Stieltjes transform of distribution G is

$$m_G(z) = \int \frac{dG(\lambda)}{\lambda - z}\,, z \in \mathbb{C}^+ (\mathsf{Im}(z) > 0)$$

Interlude: Stieltjes transforms of distributions

Stieltjes transform of distribution G is

$$m_G(z) = \int \frac{dG(\lambda)}{\lambda - z}\,, z \in \mathbb{C}^+ (\text{Im}(z) > 0)$$

For spectral distribution of sample covariance matrix,

$$m_{F_p}(z) = \frac{1}{p}\sum_{i=1}^{p} \frac{1}{l_i - z} = \frac{1}{p}\text{trace}((X'X/n - z\mathrm{Id}_p)^{-1})$$

Interlude: Stieltjes transforms of distributions

Stieltjes transform of distribution G is

$$m_G(z) = \int \frac{dG(\lambda)}{\lambda - z}\, , z \in \mathbb{C}^+ (\mathsf{Im}(z) > 0)$$

For spectral distribution of sample covariance matrix,

$$m_{F_p}(z) = \frac{1}{p} \sum_{i=1}^{p} \frac{1}{l_i - z} = \frac{1}{p} \mathsf{trace}((X'X/n - z\mathrm{Id}_p)^{-1})$$

For technical convenience, will use

$$v_{F_p}(z) = (1 - p/n)\frac{-1}{z} + \frac{p}{n} m_{F_p}(z)$$

The Marčenko-Pastur equation: characterization of limiting spectrum

Suppose Y has i.i.d entries, mean 0, sd 1, 4th absolute moment. Let $\boldsymbol{X} = \boldsymbol{Y}\boldsymbol{\Sigma}_{\boldsymbol{p}}^{\boldsymbol{1/2}}$.

The Marčenko-Pastur equation: characterization of limiting spectrum

Suppose Y has i.i.d entries, mean 0, sd 1, 4th absolute moment. Let $\boldsymbol{X} = \boldsymbol{Y}\boldsymbol{\Sigma}_{\boldsymbol{p}}^{\boldsymbol{1/2}}$.
H_p :spectral distribution of Σ_p. Assume $H_p \Longrightarrow H_\infty$.

The Marčenko-Pastur equation: characterization of limiting spectrum

Suppose Y has i.i.d entries, mean 0, sd 1, 4th absolute moment. Let $\boldsymbol{X = Y\Sigma_p^{1/2}}$.
H_p :spectral distribution of Σ_p. Assume $H_p \Longrightarrow H_\infty$.
Σ_p diagonal and $p/n \to \rho \in (0, \infty)$

Theorem 1. [Marčenko-Pastur , '67] *Let F_p be spectral distribution of X^*X/n. As $n \to \infty$,*

$$F_p \Longrightarrow F \text{ in probability}$$

The Marčenko-Pastur equation: characterization of limiting spectrum

Suppose Y has i.i.d entries, mean 0, sd 1, 4th absolute moment. Let $\boldsymbol{X = Y\Sigma_p^{1/2}}$.
H_p :spectral distribution of Σ_p. Assume $H_p \Longrightarrow H_\infty$.
Σ_p diagonal and $p/n \to \rho \in (0, \infty)$

Theorem 1. [Marčenko-Pastur , '67] *Let F_p be spectral distribution of X^*X/n. As $n \to \infty$,*

$$F_p \Longrightarrow F \text{ in probability}$$

Moreover,

$$-\frac{1}{v_F(z)} = z - \rho \int \frac{\lambda dH_\infty(\lambda)}{1 + \lambda v_F(z)} \, , \forall z \in \mathbb{C}^+$$

Marčenko-Pastur equation: precision and further developments

Marčenko-Pastur , Mat. Sbornik, '67: actually some dependence in Y allowed.
Wachter, AOP '78: a.s convergence, $2+\epsilon$ absolute moment uniformly bounded
Bai and Silverstein, JMA '95: a.s convergence, 2 moment and convergence of spectral distribution
Silverstein, JMA '95: non-diagonal covariance

Now speed of convergence and even some CLTs for linear functionals of eigenvalues... (Silverstein and Bai (AOP '04))

Spectrum shrinkage with Marčenko-Pastur equation

Question: can we use Marčenko-Pastur equation to better estimate eigenvalues?

- High-Dimensional setting: estimate spectral distribution and not individual eigenvalues

Spectrum shrinkage with Marčenko-Pastur equation

Question: can we use Marčenko-Pastur equation to better estimate eigenvalues?

- High-Dimensional setting: estimate spectral distribution and not individual eigenvalues
- Invert Marčenko-Pastur equation to do so

Spectrum shrinkage with Marčenko-Pastur equation

Question: can we use Marčenko-Pastur equation to better estimate eigenvalues?

- High-Dimensional setting: estimate spectral distribution and not individual eigenvalues
- Invert Marčenko-Pastur equation to do so
- Need fast, accurate, statistically usable solution

Spectrum shrinkage with Marčenko-Pastur equation

Question: can we use Marčenko-Pastur equation to better estimate eigenvalues?

- High-Dimensional setting: estimate spectral distribution and not individual eigenvalues
- Invert Marčenko-Pastur equation to do so
- Need fast, accurate, statistically usable solution

Proposed solution: cast the problem as convex optimization problem

Interpolation problem: observe $(z_j, v_{F_p}(z_j))_{j=1}^K$ and find estimate of H_p.

Spectrum shrinkage with Marčenko-Pastur equation

Question: can we use Marčenko-Pastur equation to better estimate eigenvalues?

- High-Dimensional setting: estimate spectral distribution and not individual eigenvalues
- Invert Marčenko-Pastur equation to do so
- Need fast, accurate, statistically usable solution

Proposed solution: cast the problem as convex optimization problem

Interpolation problem: observe $(z_j, v_{F_p}(z_j))_{j=1}^{K}$ and find estimate of H_p.

Fairly natural idea to use Marčenko-Pastur equation. Only problem is how to invert it. (See also work by Burda and al.)

Numerical inversion of Marčenko-Pastur equation: discrete version

Recall: $(z_j, v_{F_p}(z_j))_{j=1}^{K}$ observable.
Aim: find $\widehat{H}_p$ such that

$$-\frac{1}{v_{F_p}(z_j)} \simeq z_j - \frac{p}{n}\int \frac{\lambda}{1+\lambda v_{F_p}(z_j)} d\widehat{H}_p(\lambda)\ , \ \ \forall z_j,\ j=1,\ldots,K$$

Numerical inversion of Marčenko-Pastur equation: discrete version

Recall: $(z_j, v_{F_p}(z_j))_{j=1}^K$ observable.
Aim: find $\widehat{H}_p$ such that

$$-\frac{1}{v_{F_p}(z_j)} \simeq z_j - \frac{p}{n} \int \frac{\lambda}{1 + \lambda v_{F_p}(z_j)} d\widehat{H}_p(\lambda) \ , \ \ \forall z_j, \, j = 1, \ldots, K$$

Strategy: make the problem discrete

- Choose grid $(t_1, \ldots, t_N)$ for the λ's
- Approximate dH_p by $\sum w_i \delta_{t_i}$

Numerical inversion of Marčenko-Pastur equation: discrete version

Recall: $(z_j, v_{F_p}(z_j))_{j=1}^K$ observable.
Aim: find $\widehat{H}_p$ such that

$$-\frac{1}{v_{F_p}(z_j)} \simeq z_j - \frac{p}{n}\int \frac{\lambda}{1+\lambda v_{F_p}(z_j)} d\widehat{H}_p(\lambda)\ , \ \ \forall z_j,\, j=1,\ldots,K$$

Strategy: make the problem discrete

- Choose grid $(t_1, \ldots, t_N)$ for the λ's
- Approximate dH_p by $\sum w_i \delta_{t_i}$, hence

$$\int \frac{\lambda}{1+\lambda v_{F_p}(z)} dH_p(\lambda) \simeq \sum_{i=1}^N w_i \frac{t_i}{1+t_i v_{F_p}(z)}$$

- Constraints: $\sum_{i=1}^N w_i = 1$ and $w_i \geq 0, \forall i$

Possible formulation of Marčenko-Pastur shrinkage

Our problem becomes find "best" $(w_1, \ldots, w_N)$. This can be made into **convex program**.

Call $g(z_j) = (z_j + 1/v_{F_p}(z_j))n/p$.

Possible formulation:

$$\min_{(w_1,\ldots,w_N)} \max_{j=1,\ldots,K} \left|\text{Re}\left(g(z_j) - \sum_{i=1}^{N} \frac{w_i t_i}{1 + t_i v_{F_p}(z_j)}\right)\right|, \left|\text{Im}\left(g(z_j) - \sum_{i=1}^{N} \frac{w_i t_i}{1 + t_i v_{F_p}(z_j)}\right)\right|$$

$$\text{subject to } \sum_{i=1}^{N} w_i = 1 \text{ and } w_i \geq 0, \ \forall i$$

Possible formulation of Marčenko-Pastur shrinkage

Our problem becomes find "best" $(w_1, \ldots, w_N)$. This can be made into **convex program**.

Call $g(z_j) = (z_j + 1/v_{F_p}(z_j))n/p$.

Possible formulation:

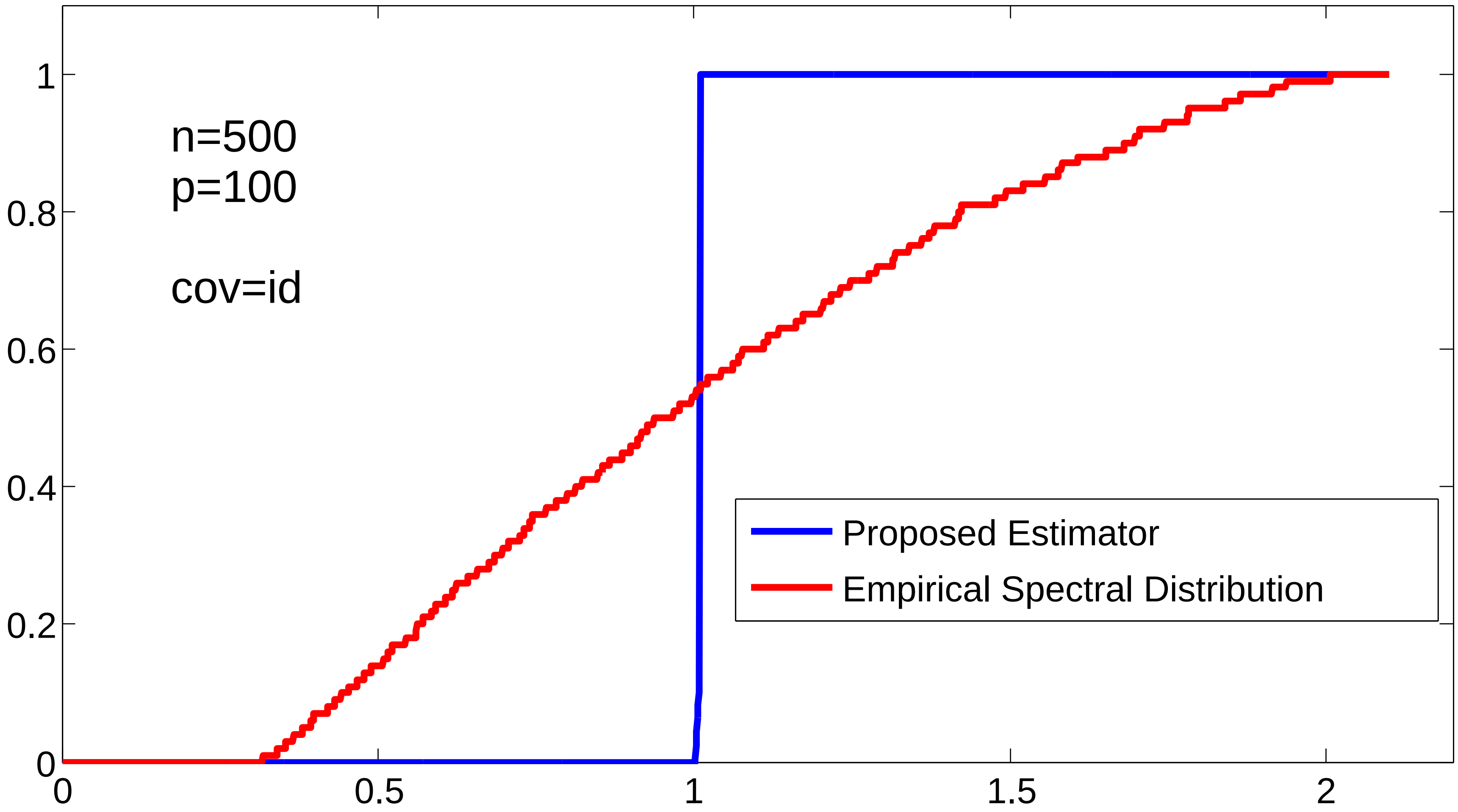
$$\min_{(w_1,\ldots,w_N)} \max_{j=1,\ldots,K} \left|\text{Re}\left(g(z_j) - \sum_{i=1}^{N} \frac{w_i t_i}{1 + t_i v_{F_p}(z_j)}\right)\right|, \left|\text{Im}\left(g(z_j) - \sum_{i=1}^{N} \frac{w_i t_i}{1 + t_i v_{F_p}(z_j)}\right)\right|$$

$$\text{subject to } \sum_{i=1}^{N} w_i = 1 \text{ and } w_i \geq 0, \ \forall i$$

This is a linear program (LP)

Many other possibilities also convex optimization problems Solution of optimization problem gives **estimate of CDF**. Estimated eigenvalues: **quantiles** of this CDF.

Example 1: population covariance = Id
Population Spectrum = δ_1



Example 1: population covariance = Id
Population Spectrum = δ_1

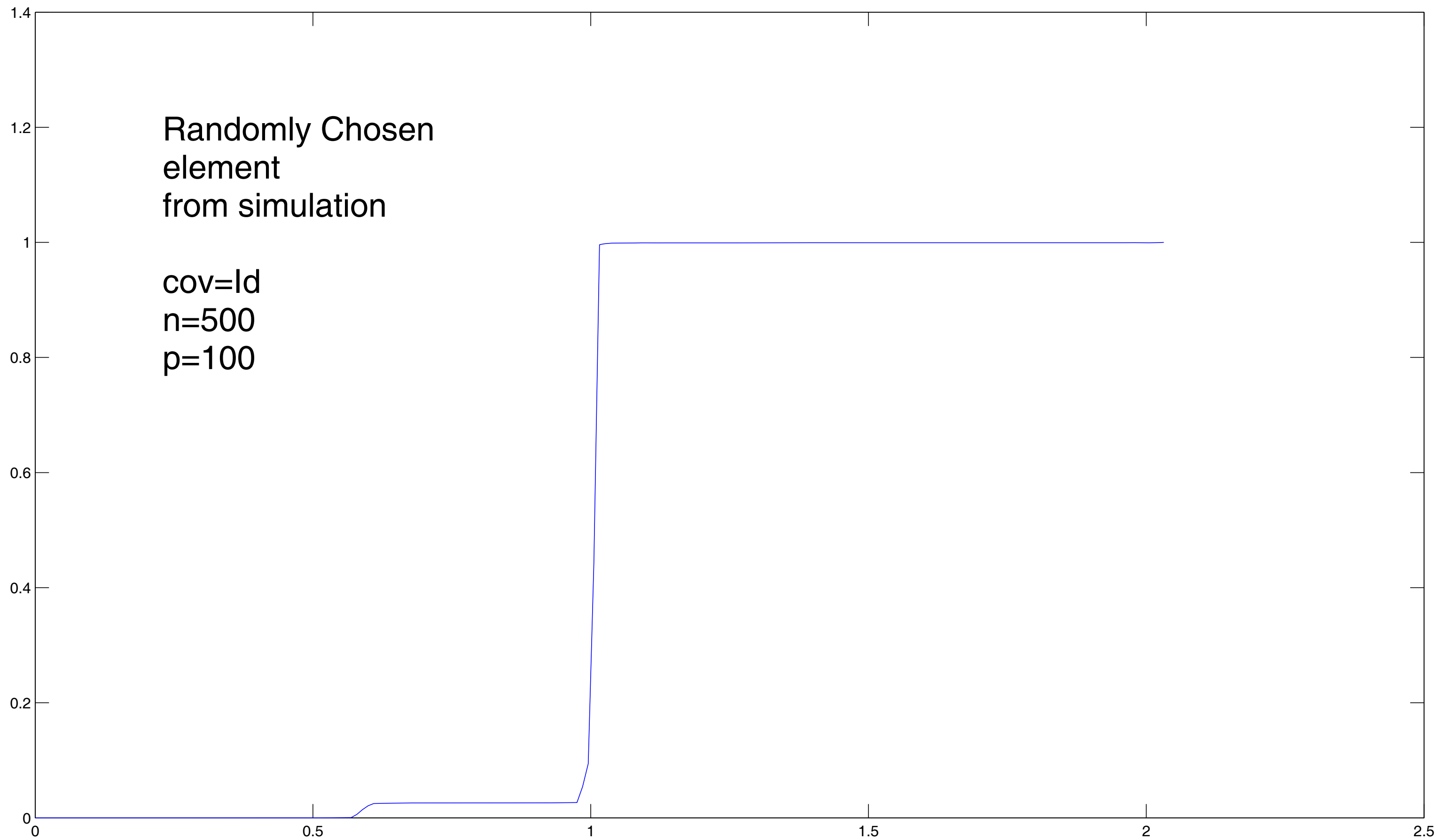
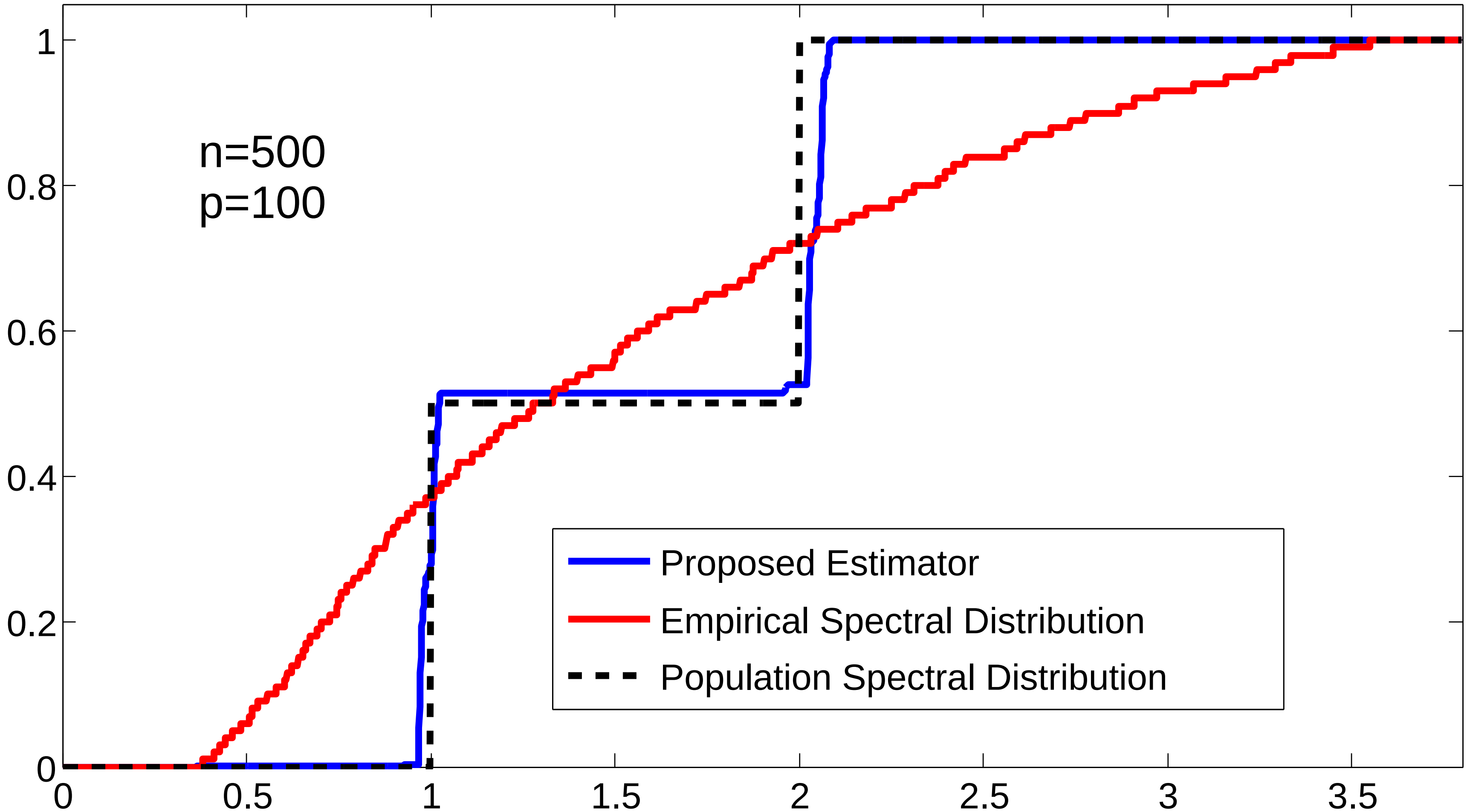
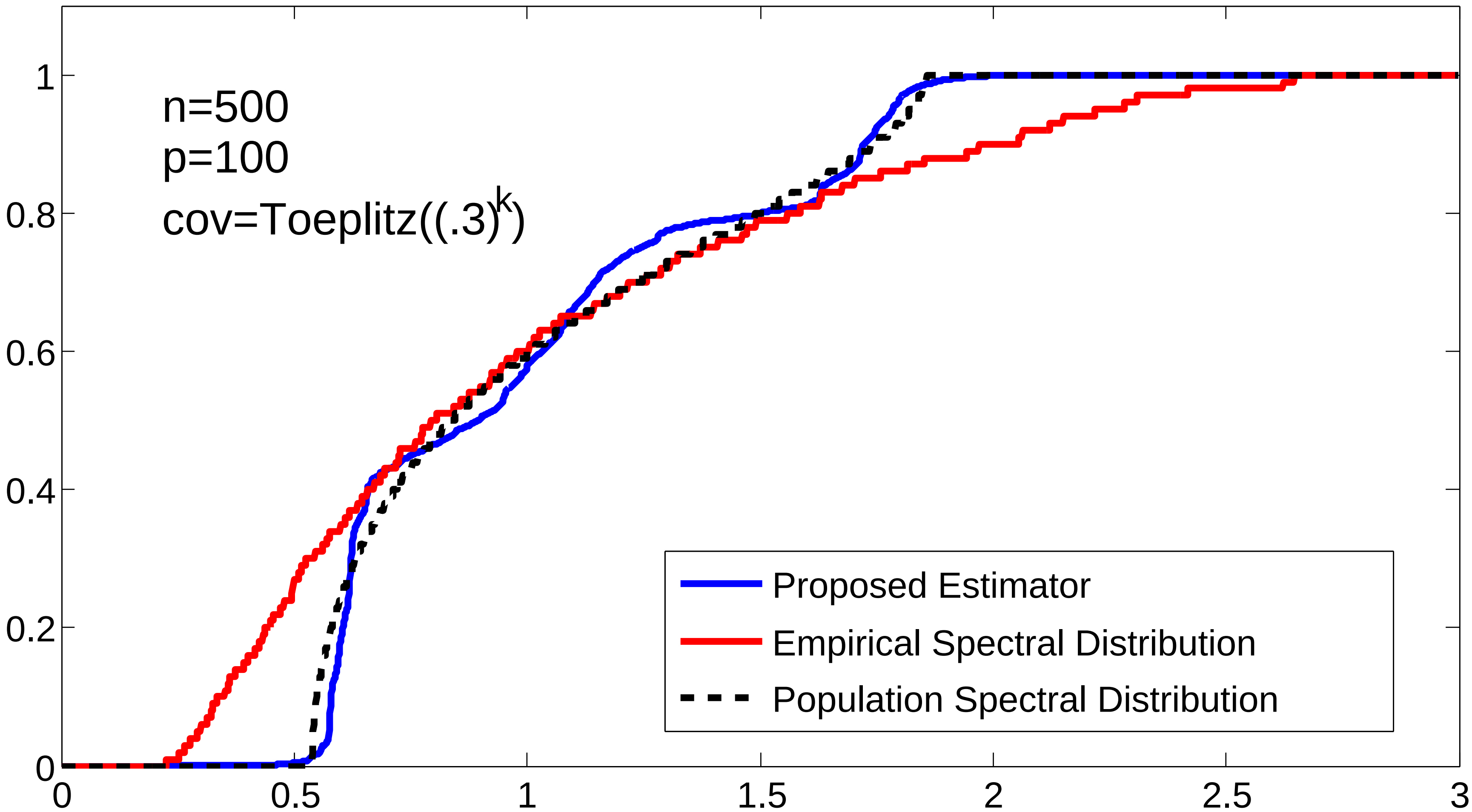


Figure 1: Example of Estimator

Example 2: population covariance = diagonal(50 1's, 50 2's)
Population Spectrum = $1/2\delta_1 + 1/2\delta_2$



Example 3: population covariance = Toeplitz($.3^k$), $k = 0$ to 100



Part II : Testing and recent theoretical results

Limit of largest eigenvalue: case of $\Sigma_p = \mathrm{Id}_p$

Consider complex Gaussian case $X_i \sim \mathcal{N}_{\mathbb{C}}(0, \Sigma_p) = \mathcal{N}(0, \Sigma_p/2) + i\mathcal{N}(0, \Sigma_p/2)$

In case $\Sigma_p = \mathrm{Id}_p$, results of Forrester, Johansson, Johnstone show, if $p/n \to \gamma \in (0, \infty)$

$$n^{2/3} \frac{l_1(X^* X/n) - \mu_{n,p}}{\sigma_{n,p}} \Longrightarrow \mathsf{TW}_2$$

Limit of largest eigenvalue: case of $\Sigma_p = \mathrm{Id}_p$

Consider complex Gaussian case $X_i \sim \mathcal{N}_{\mathbb{C}}(0, \Sigma_p) = \mathcal{N}(0, \Sigma_p/2) + i\mathcal{N}(0, \Sigma_p/2)$

In case $\Sigma_p = \mathrm{Id}_p$, results of Forrester, Johansson, Johnstone show, if $p/n \to \gamma \in (0, \infty)$

$$n^{2/3} \frac{l_1(X^*X/n) - \mu_{n,p}}{\sigma_{n,p}} \Longrightarrow \mathsf{TW}_2$$

where

$$\mu_{n,p} = (1 + \sqrt{p/n})^2$$
$$\sigma_{n,p} = (1 + \sqrt{p/n}) \left(1 + \sqrt{\frac{n}{p}}\right)^{1/3}$$

Limit of largest eigenvalue: case of $\Sigma_p = \mathrm{Id}_p$

Consider complex Gaussian case $X_i \sim \mathcal{N}_{\mathbb{C}}(0, \Sigma_p) = \mathcal{N}(0, \Sigma_p/2) + i\mathcal{N}(0, \Sigma_p/2)$

In case $\Sigma_p = \mathrm{Id}_p$, results of Forrester, Johansson, Johnstone show, if $p/n \to \gamma \in (0, \infty)$

$$n^{2/3} \frac{l_1(X^*X/n) - \mu_{n,p}}{\sigma_{n,p}} \Longrightarrow \mathrm{TW}_2$$

where

$$\mu_{n,p} = (1 + \sqrt{p/n})^2$$
$$\sigma_{n,p} = (1 + \sqrt{p/n}) \left(1 + \sqrt{\frac{n}{p}}\right)^{1/3}$$

Results valid $\gamma = 0$ or ∞ (NEK)

Beyond Id case and a rate result

Theorem 2. [NEK] *Let* X *be* $n \times p$ *matrix.* $X_i \overset{iid}{\backsim} \mathcal{N}_{\mathbb{C}}(0, \Sigma_p)$. *Let* H_p *be spectral distribution of* Σ_p. *Suppose* $n \geq p$ *and* n/p *bounded. Call*

$$c = c(\Sigma_p, n, p)\,,\; c \in [0, 1/\lambda_1(\Sigma_p))\;\; : \int \left(\frac{\lambda c}{1 - \lambda c}\right)^2 dH_p(\lambda) = \frac{n}{p}$$

Then

$$n^{2/3} \frac{l_1(X^* X/n) - \mu}{\sigma} \Longrightarrow \mathit{TW}_2\,,$$

Beyond Id case and a rate result

Theorem 2. [NEK] *Let* X *be* $n \times p$ *matrix.* $X_i \overset{iid}{\backsim} \mathcal{N}_{\mathbb{C}}(0, \Sigma_p)$. *Let* H_p *be spectral distribution of* Σ_p. *Suppose* $n \geq p$ *and* n/p *bounded. Call*

$$c = c(\Sigma_p, n, p)\,,\; c \in [0, 1/\lambda_1(\Sigma_p))\; : \int \left(\frac{\lambda c}{1 - \lambda c}\right)^2 dH_p(\lambda) = \frac{n}{p}$$

Then

$$n^{2/3} \frac{l_1(X^* X/n) - \mu}{\sigma} \Longrightarrow \textit{TW}_2\,,$$

provided $\limsup \lambda_1(\Sigma_p) < \infty$, $\limsup \lambda_1(\Sigma_p) c(\Sigma_p, n, p) < 1$, $\liminf \lambda_p(\Sigma_p) > 0$

Beyond Id case and a rate result

Theorem 2. [NEK] *Let* X *be* $n \times p$ *matrix.* $X_i \overset{iid}{\backsim} \mathcal{N}_{\mathbb{C}}(0, \Sigma_p)$. *Let* H_p *be spectral distribution of* Σ_p. *Suppose* $n \geq p$ *and* n/p *bounded. Call*

$$c = c(\Sigma_p, n, p)\,,\; c \in [0, 1/\lambda_1(\Sigma_p))\;:\; \int \left(\frac{\lambda c}{1 - \lambda c}\right)^2 dH_p(\lambda) = \frac{n}{p}$$

Then

$$n^{2/3} \frac{l_1(X^* X/n) - \mu}{\sigma} \Longrightarrow \mathit{TW}_2\,,$$

provided $\limsup \lambda_1(\Sigma_p) < \infty$, $\limsup \lambda_1(\Sigma_p) c(\Sigma_p, n, p) < 1$, $\liminf \lambda_p(\Sigma_p) > 0$

Theorem 3. [NEK] $\boldsymbol{\rho \in (0, 1]}$ *and* $\Sigma_p = \mathrm{Id}_p$. $\exists\; \tilde{\mu}_{n,p}$, $\tilde{\sigma}_{n,p}$, *and function* C, *continuous, non-increasing; as* $n, p \to \infty$, *and* $n/p \to \rho \in \mathbb{R}_+^*$, $\forall s_0$, $\exists N(s_0, \rho)$,

$$\forall s \geq s_0,\; \textit{if } (n \wedge p) \geq N(s_0, \rho),$$

$$(n \wedge p)^{2/3} \left| P\left(\frac{l_1 - \tilde{\mu}_{n,p}}{\tilde{\sigma}_{n,p}} \leq s\right) - F_2(s) \right| \leq C(s_0) \exp(-s)$$

Non Id Covariance problem

- If using convergence results for testing, need to know power
- In applications, most often $\Sigma_p \neq \mathrm{Id}$
- Influence of small eigenvalues on the largest empirical eigenvalue?
- Do we still get Tracy-Widom limit, or is it artifact of $\Sigma_p = \mathrm{Id}$?

Non Id Covariance problem

- If using convergence results for testing, need to know power
- In applications, most often $\Sigma_p \neq \mathrm{Id}$
- Influence of small eigenvalues on the largest empirical eigenvalue?
- Do we still get Tracy-Widom limit, or is it artifact of $\Sigma_p = \mathrm{Id}$?

Important recent work ('04) by **Baik, Ben Arous and Péché** in Gaussian complex case when Σ_p finite perturbation of Id_p. (k eigenvalues different from 1, all others equal to 1.)

Non Id covariance: fairly general result for complex case

Theorem. [NEK] *Let X be $n \times p$ matrix. $X_i \overset{iid}{\backsim} \mathcal{N}_{\mathbb{C}}(0, \Sigma_p)$. Let H_p be spectral distribution of Σ_p. Suppose $n \geq p$ and n/p bounded. Call*

$$c = c(\Sigma_p, n, p)\,,\; c \in [0, 1/\lambda_1(\Sigma_p)) \;:\; \int \left(\frac{\lambda c}{1 - \lambda c}\right)^2 dH_p(\lambda) = \frac{n}{p}$$

Non Id covariance: fairly general result for complex case

Theorem. [NEK] *Let* X *be* $n \times p$ *matrix.* $X_i \overset{iid}{\backsim} \mathcal{N}_{\mathbb{C}}(0, \Sigma_p)$. *Let* H_p *be spectral distribution of* Σ_p. *Suppose* $n \geq p$ *and* n/p *bounded. Call*

$$c = c(\Sigma_p, n, p)\,,\; c \in [0, 1/\lambda_1(\Sigma_p)) \;:\; \int \left(\frac{\lambda c}{1 - \lambda c}\right)^2 dH_p(\lambda) = \frac{n}{p}$$

Let

$$\boldsymbol{\mu} = \frac{1}{c}\left(1 + \frac{p}{n}\int \frac{\lambda c}{1 - \lambda c} dH_p(\lambda)\right) \;,\; \boldsymbol{\sigma}^{\mathbf{3}} = \frac{1}{c^3}\left(1 + \frac{p}{n}\int \left(\frac{\lambda c}{1 - \lambda c}\right)^3 dH_p(\lambda)\right)$$

Then

$$n^{2/3}\frac{l_1(X^*X/n) - \mu}{\sigma} \Longrightarrow TW_2\,,$$

Non Id covariance: fairly general result for complex case

Theorem. [NEK] *Let* X *be* $n \times p$ *matrix.* $X_i \overset{iid}{\backsim} \mathcal{N}_{\mathbb{C}}(0, \Sigma_p)$. *Let* H_p *be spectral distribution of* Σ_p. *Suppose* $n \geq p$ *and* n/p *bounded. Call*

$$c = c(\Sigma_p, n, p)\,,\ c \in [0, 1/\lambda_1(\Sigma_p))\ :\int \left(\frac{\lambda c}{1-\lambda c}\right)^2 dH_p(\lambda) = \frac{n}{p}$$

Let

$$\boldsymbol{\mu} = \frac{1}{c}\left(1 + \frac{p}{n}\int \frac{\lambda c}{1-\lambda c} dH_p(\lambda)\right)\ ,\ \boldsymbol{\sigma^3} = \frac{1}{c^3}\left(1 + \frac{p}{n}\int \left(\frac{\lambda c}{1-\lambda c}\right)^3 dH_p(\lambda)\right)$$

Then

$$n^{2/3}\frac{l_1(X^*X/n) - \mu}{\sigma} \Longrightarrow \mathit{TW}_2\ ,$$

provided $\limsup \lambda_1(\Sigma_p) < \infty$, $\limsup \lambda_1(\Sigma_p) c(\Sigma_p, n, p) < 1$, $\liminf \lambda_p(\Sigma_p) > 0$

Remarks about previous theorem

- $c(\Sigma_p, n, p)$ easily computed for given model
- $l_1(X^*X)/n$ inconsistent estimator of λ_1.
- Bias > 0 (Jensen) and increasing with p/n.
- Condition $\boldsymbol{\limsup \lambda_1(\Sigma_p) c(\Sigma_p, n, p) < 1}$ extremely important; failure may lead to other limiting distribution, centering and scaling sequences
- Condition $\liminf \lambda_p(\Sigma_p) > 0$ can probably be weakened and result still holds

Practical examples for which conditions hold

- Fixed spectral distribution asymptotics
- Equally spaced eigenvalues on $[a, b]$, $a > 0$, $b < \infty$
- Finite order Toeplitz matrices, when eigenvalues bounded away from 0
- In general Toeplitz matrices for which

$$f(\omega) = 1 + 2\sum_{k=1}^{\infty} a_k \cos(k\omega)$$

is not too "wild", $\inf_{[0,2\pi]} f(\omega) > 0$ and $\sup_{[0,2\pi]} f(\omega) < \infty$

Idea of proof

- Joint distribution of eigenvalues of X^*X is known

- Write $P(l_1 \leq \mu_{n,p} + \sigma_{n,p}s)$ as determinant of $(\mathrm{Id}-)$ trace class operator on $L^2[s,\infty)$

- Operator is product of two Hilbert-Schmidt operators, with kernels $K(x,y) = K(x+y-s)$

Idea of proof

- Joint distribution of eigenvalues of X^*X is known

- Write $P(l_1 \leq \mu_{n,p} + \sigma_{n,p} s)$ as determinant of $(\mathrm{Id} -)$ trace class operator on $L^2[s, \infty)$

- Operator is product of two Hilbert-Schmidt operators, with kernels $K(x, y) = K(x + y - s)$

- Representation of kernels as contour integrals

- Do "steepest" descent analysis to show closeness of kernels to Airy

Idea of proof

- Joint distribution of eigenvalues of X^*X is known
- Write $P(l_1 \leq \mu_{n,p} + \sigma_{n,p} s)$ as determinant of $(\mathrm{Id}-)$ trace class operator on $L^2[s, \infty)$
- Operator is product of two Hilbert-Schmidt operators, with kernels $K(x, y) = K(x + y - s)$
- Representation of kernels as contour integrals
- Do "steepest" descent analysis to show closeness of kernels to Airy
- Difficulty: find c and type of limit. Then find path.
- Intuition for c comes from realizing connection with Marčenko-Pastur equation

Elementary manipulations

Using result in Baik, Ben Arous, Péché, and simple manipulations, can show

$$P\left(\frac{l_1}{n} \leq \mu_{n,p} + \sigma_{n,p} s\right) = \det(I - S_{n,p}|_{L^2[s,\infty)})$$

where

$$S_{n,p} = A_{n,p} B_{n,p}$$

Elementary manipulations

Using result in Baik, Ben Arous, Péché, and simple manipulations, can show

$$P\left(\frac{l_1}{n} \leq \mu_{n,p} + \sigma_{n,p} s\right) = \det(I - S_{n,p}|_{L^2[s,\infty)})$$

where

$$S_{n,p} = A_{n,p} B_{n,p}$$

$A_{n,p}$ and $B_{n,p}$ operators on $L^2([s,\infty))$ with kernels $A_{n,p}(x,y) = A_{n,p}(x+y-s)$ and similarly $B_{n,p}$. Moreover,

$$A_{n,p}(x) = -\frac{n\sigma_{n,p}}{2\pi i}\int_\Gamma \mathrm{e}^{-n\sigma_{n,p}x(z-q)}\mathrm{e}^{-n\mu_{n,p}(z-q)} z^n \prod_{k=1}^{p} \frac{1}{1-z\lambda_k} dz \text{ and}$$

$$B_{n,p}(x) = \frac{n\sigma_{n,p}}{2\pi i}\int_\Xi \mathrm{e}^{n\sigma_{n,p}x(z-q)}\mathrm{e}^{n\mu_{n,p}(z-q)} \frac{1}{z^n} \prod_{k=1}^{p} (1-\lambda_k z) dz \ .$$

for q a regularization parameter to be chosen later

Toward a connection with Marčenko-Pastur equation

Integrand in definition of $A_{n,p}$ can be written

$$e^{-n\sigma_{n,p}x(z-q)}e^{nf(z)} \text{ with}$$

$$f(z) = -\mu_{n,p}(z-q) + \log(z) - \frac{1}{n}\sum_{k=1}^{p}\log(1-z\lambda_k)\text{ , or}$$

$$f(z) = -\mu_{n,p}(z-q) + \log(z) - \frac{p}{n}\int \log(1-z\lambda)dH_p(\lambda)\ .$$

Now

$$\begin{aligned} f'(z) &= -\mu_{n,p} + \frac{1}{z} + \frac{p}{n}\int \frac{\lambda}{1-z\lambda}dH_p(\lambda) \\ &= -\mu_{n,p} + g_p(-z) \end{aligned}$$

Consequences of Marčenko-Pastur connection

Interpretation of result in Marčenko-Pastur '67 (also in Silverstein and Choi, JMA '95) : "essentially", support of F_∞ determined by $g_\infty(x_0)$ when $g'_\infty(x_0) = 0$.

$$f''(z) = -g'_p(-z)$$

so these x_0 "should" be saddle points of order 2 of f.

Consequences of Marčenko-Pastur connection

Interpretation of result in Marčenko-Pastur '67 (also in Silverstein and Choi, JMA '95) : "essentially", support of F_∞ determined by $g_\infty(x_0)$ when $g'_\infty(x_0) = 0$.

$$f''(z) = -g'_p(-z)$$

so these x_0 "should" be saddle points of order 2 of f.
Will see this is very important for Tracy-Widom limits

Consequences of Marčenko-Pastur connection

Interpretation of result in Marčenko-Pastur '67 (also in Silverstein and Choi, JMA '95) : "essentially", support of F_∞ determined by $g_\infty(x_0)$ when $g'_\infty(x_0) = 0$.

$$f''(z) = -g'_p(-z)$$

so these x_0 "should" be saddle points of order 2 of f.
Will see this is very important for Tracy-Widom limits

Now $f''(z_0) = 0$ is equivalent to

$$\int \left(\frac{\lambda z_0}{1 - \lambda z_0}\right)^2 dH_p(\lambda) = \frac{n}{p} \; .$$

Source of $c(\Sigma_p, n, p)$ and choices of $\mu_{n,p}$ and $\sigma_{n,p}$

Consequences of heuristic analysis

Heuristic: if $c = z_0 = -x_0$, $f'(c) = f''(c) = 0$,
so $f(z) \simeq f(c) + f^{(3)}(c)(z - c)^3/6$

Consequences of heuristic analysis

Heuristic: if $c = z_0 = -x_0$, $f'(c) = f''(c) = 0$,
so $f(z) \simeq f(c) + f^{(3)}(c)(z-c)^3/6$

$$A_{n,p}(x) \simeq -\frac{n\sigma_{n,p}}{2\pi i} \mathrm{e}^{-n\sigma_{n,p}x(c-q)} \mathrm{e}^{nf(c)} \int_\Gamma \exp\left(-xn\sigma_{n,p}(z-c) + nf^{(3)}(c)\frac{(z-c)^3}{6}\right) dz$$

Choose $(n\sigma_{n,p})^3 = nf^{(3)}(c)/2$, so if $a = \tau(z) = n\sigma_{n,p}(z-c)$,

$$A_{n,p}(x) \simeq -\frac{\mathrm{e}^{nf(c)}}{2\pi i} \mathrm{e}^{-n\sigma_{n,p}x(c-q)} \int_{\tau(\Gamma)} \exp\left(-xa + \frac{a^3}{3}\right) da\ .$$

Consequences of heuristic analysis

Heuristic: if $c = z_0 = -x_0$, $f'(c) = f''(c) = 0$,
so $f(z) \simeq f(c) + f^{(3)}(c)(z-c)^3/6$

$$A_{n,p}(x) \simeq -\frac{n\sigma_{n,p}}{2\pi i} \mathrm{e}^{-n\sigma_{n,p}x(c-q)} \mathrm{e}^{nf(c)} \int_\Gamma \exp\left(-xn\sigma_{n,p}(z-c) + nf^{(3)}(c)\frac{(z-c)^3}{6}\right) dz$$

Choose $(n\sigma_{n,p})^3 = nf^{(3)}(c)/2$, so if $a = \tau(z) = n\sigma_{n,p}(z-c)$,

$$A_{n,p}(x) \simeq -\frac{\mathrm{e}^{nf(c)}}{2\pi i} \mathrm{e}^{-n\sigma_{n,p}x(c-q)} \int_{\tau(\Gamma)} \exp\left(-xa + \frac{a^3}{3}\right) da \ .$$

But, for "good" contour $\mathcal{L}$,

$$\mathrm{Ai}(x) = \frac{1}{2\pi i} \int_{\mathcal{L}} \exp\left(-xv + \frac{v^3}{3}\right) dv \ .$$

Consequences of heuristic analysis

Heuristic: if $c = z_0 = -x_0$, $f'(c) = f''(c) = 0$,
so $f(z) \simeq f(c) + f^{(3)}(c)(z-c)^3/6$

$$A_{n,p}(x) \simeq -\frac{n\sigma_{n,p}}{2\pi i} \mathrm{e}^{-n\sigma_{n,p}x(c-q)} \mathrm{e}^{nf(c)} \int_\Gamma \exp\left(-xn\sigma_{n,p}(z-c) + nf^{(3)}(c)\frac{(z-c)^3}{6}\right) dz$$

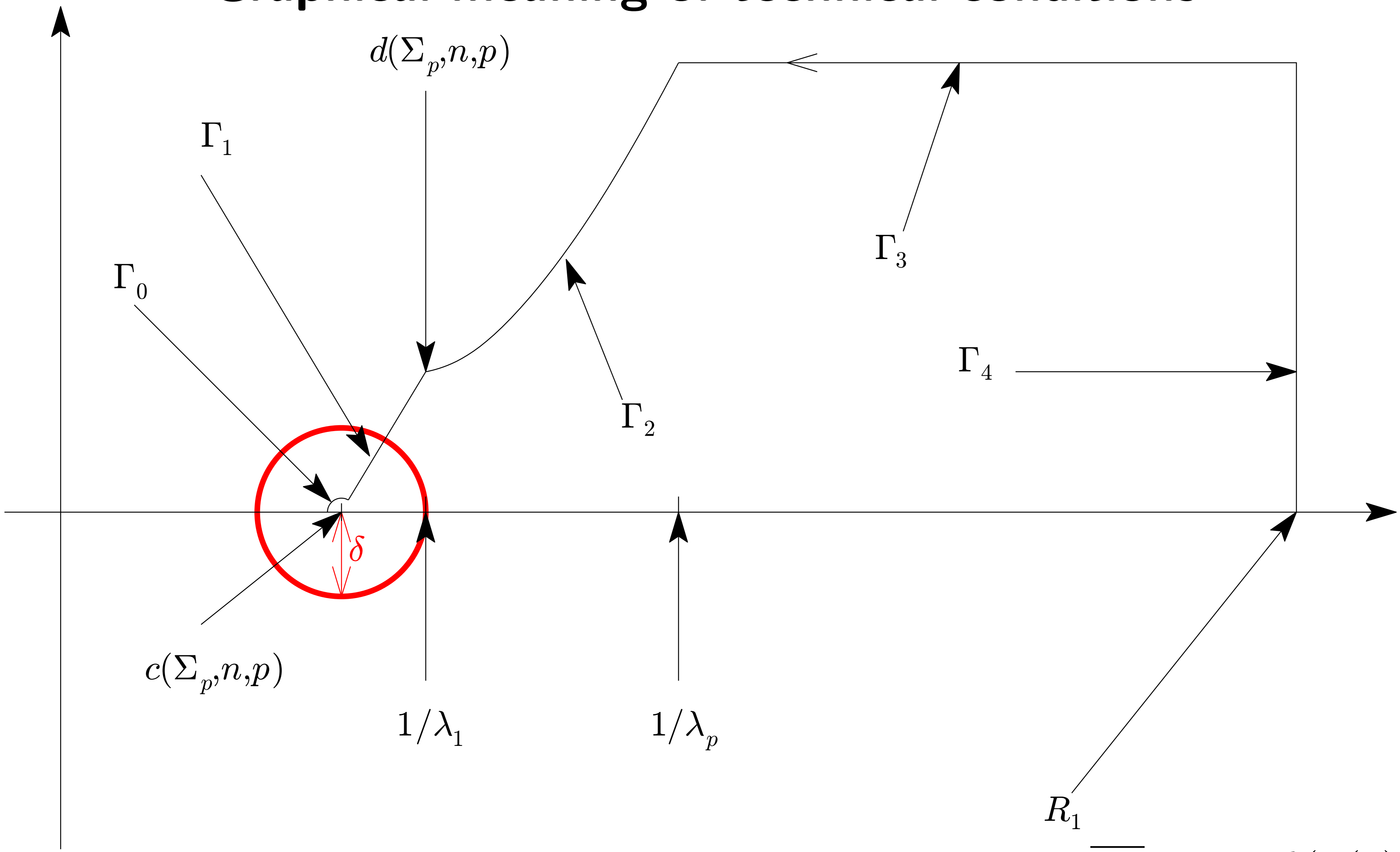
Choose $(n\sigma_{n,p})^3 = nf^{(3)}(c)/2$, so if $a = \tau(z) = n\sigma_{n,p}(z-c)$,

$$A_{n,p}(x) \simeq -\frac{\mathrm{e}^{nf(c)}}{2\pi i} \mathrm{e}^{-n\sigma_{n,p}x(c-q)} \int_{\tau(\Gamma)} \exp\left(-xa + \frac{a^3}{3}\right) da \ .$$

But, for "good" contour $\mathcal{L}$,

$$\mathrm{Ai}(x) = \frac{1}{2\pi i} \int_{\mathcal{L}} \exp\left(-xv + \frac{v^3}{3}\right) dv \ .$$

Graphical meaning of technical conditions



Analysis ends up being about finding a path $\Gamma = \Gamma_+ \cup \overline{\Gamma_+}$ with $\Re(f(z))$ decreasing on most of Γ_+

Take away messages

A few statements to be made rigorous:

- From extreme eigenvalue standpoint interesting saddle points of f "have" to be saddle points of order 2
- Saddle points of order 2 lead to Airy function
- Natural scaling necessarily $n^{-2/3}$ for $\sigma_{n,p}$
- $\mu_{n,p}$ set by condition $f'(c)$

Take away messages

A few statements to be made rigorous:

- From extreme eigenvalue standpoint interesting saddle points of f "have" to be saddle points of order 2

- Saddle points of order 2 lead to Airy function

- Natural scaling necessarily $n^{-2/3}$ for $\sigma_{n,p}$

- $\mu_{n,p}$ set by condition $f'(c)$

Saddle points of **order 2** intimately linked to
Marčenko-Pastur equation

Rates of convergence and good centering and scaling

Theorem. [NEK] $\boldsymbol{\rho \in (0,1]}, \Sigma_p = \mathrm{Id}_p$. *In complex case, let* l_1 *denote largest eigenvalue of* $X^* X$. *Then,* $\exists\ \tilde{\mu}_{n,p}$, $\tilde{\sigma}_{n,p}$, *and function* C, *such that as* n, p *tend to* $+\infty$, *and* $n/p \to \rho \in \mathbb{R}_+^*$, $\forall s_0$, $\exists N(s_0, \rho)$,

$$\forall s \geq s_0,\ \textit{if}\ (n \wedge p) \geq N(s_0, \rho),$$

$$(n \wedge p)^{2/3} \left| P\left(\frac{l_1 - \tilde{\mu}_{n,p}}{\tilde{\sigma}_{n,p}} \leq s\right) - F_2(s) \right| \leq C(s_0) \exp(-s)$$

C continuous non increasing function of s
Rate analysis practically important for improving centering and scaling.

Centering and scaling sequence

With

$$\mu_{np} = (\sqrt{n-1/2} + \sqrt{p-1/2})^2 \,, \qquad (\asymp n)$$

$$\sigma_{np} = (\sqrt{n-1/2} + \sqrt{p-1/2}) \left(\frac{1}{\sqrt{n-1/2}} + \frac{1}{\sqrt{p-1/2}} \right)^{1/3} , \quad (\asymp n^{1/3})$$

$$\tilde{\mu}_{np} = \left(\frac{1}{\sigma_{p+1,n}^{1/2}} + \frac{1}{\sigma_{p,n+1}^{1/2}} \right) \left(\frac{1}{\mu_{p+1,n}\sigma_{p+1,n}^{1/2}} + \frac{1}{\mu_{p,n+1}\sigma_{p,n+1}^{1/2}} \right)^{-1}$$

$$\gamma_{n,p} = \frac{\mu_{n,p+1}\sigma_{n+1,p}^{1/2}}{\mu_{n+1,p}\sigma_{n,p+1}^{1/2}}$$

$$\tilde{\sigma}_{n,p} = (1 + \gamma_{n,p}) \left(\frac{1}{\sigma_{n,p+1}} + \frac{\gamma_{n,p}}{\sigma_{n+1,p}} \right)^{-1}$$

Summary

- Numerical aspect
  - Spectral distribution approach takes advantage of high-dimension
  - Possible to numerically invert Marčenko-Pastur equation via convex optimization
  - $p > n$ not a problem

Summary

- Numerical aspect
  - Spectral distribution approach takes advantage of high-dimension
  - Possible to numerically invert Marčenko-Pastur equation via convex optimization
  - $p > n$ not a problem
- Theoretical aspect
  - Fluctuation behavior of top eigenvalue linked with bulk results
  - Rate results in Id_p case
  - $p/n \to 0$ not a problem

Summary

- Numerical aspect
  - Spectral distribution approach takes advantage of high-dimension
  - Possible to numerically invert Marčenko-Pastur equation via convex optimization
  - $p > n$ not a problem
- Theoretical aspect
  - Fluctuation behavior of top eigenvalue linked with bulk results
  - Rate results in Id_p case
  - $p/n \to 0$ not a problem
- Results in random matrix theory potentially useful for statisticians, both practice and theory

Thanks

Organizers of the conference

Iain Johnstone

Peter Bickel(Stat, UCB), John Rice (Stat, UCB)

Laurent El Ghaoui (EECS, UCB), Alexandre d'Aspremont (ORFE, Princeton)

Toeplitz Covariance Matrices

$$\Sigma = \begin{pmatrix} a_0 & a_1 & a_2 & \dots \\ a_1 & a_0 & a_1 & \dots \\ a_2 & a_1 & a_0 & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix}$$

- No explicit formula for eigenvalues in finite dimension
- Asymptotic behavior well studied and understood
- Aside: here symmetric matrix (covariance), but need not be

Back

A couple statistical issues

1)How to measure **quality** of estimator?
Use Lévy Metric:

$$L(F,G) = \inf\{\epsilon > 0 : F(x-\epsilon) - \epsilon \leq G(x) \leq F(x+\epsilon) + \epsilon \ \ , \forall x\}$$

A couple statistical issues

1)How to measure **quality** of estimator?
Use Lévy Metric:

$$L(F,G) = \inf\{\epsilon > 0 : F(x-\epsilon) - \epsilon \leq G(x) \leq F(x+\epsilon) + \epsilon \ \ , \forall x\}$$

A few examples:
In what follows, $n = 500$, $p = 100$
1,000 independent repetitions

Example 1: population covariance = Id
Population Spectrum = δ_1

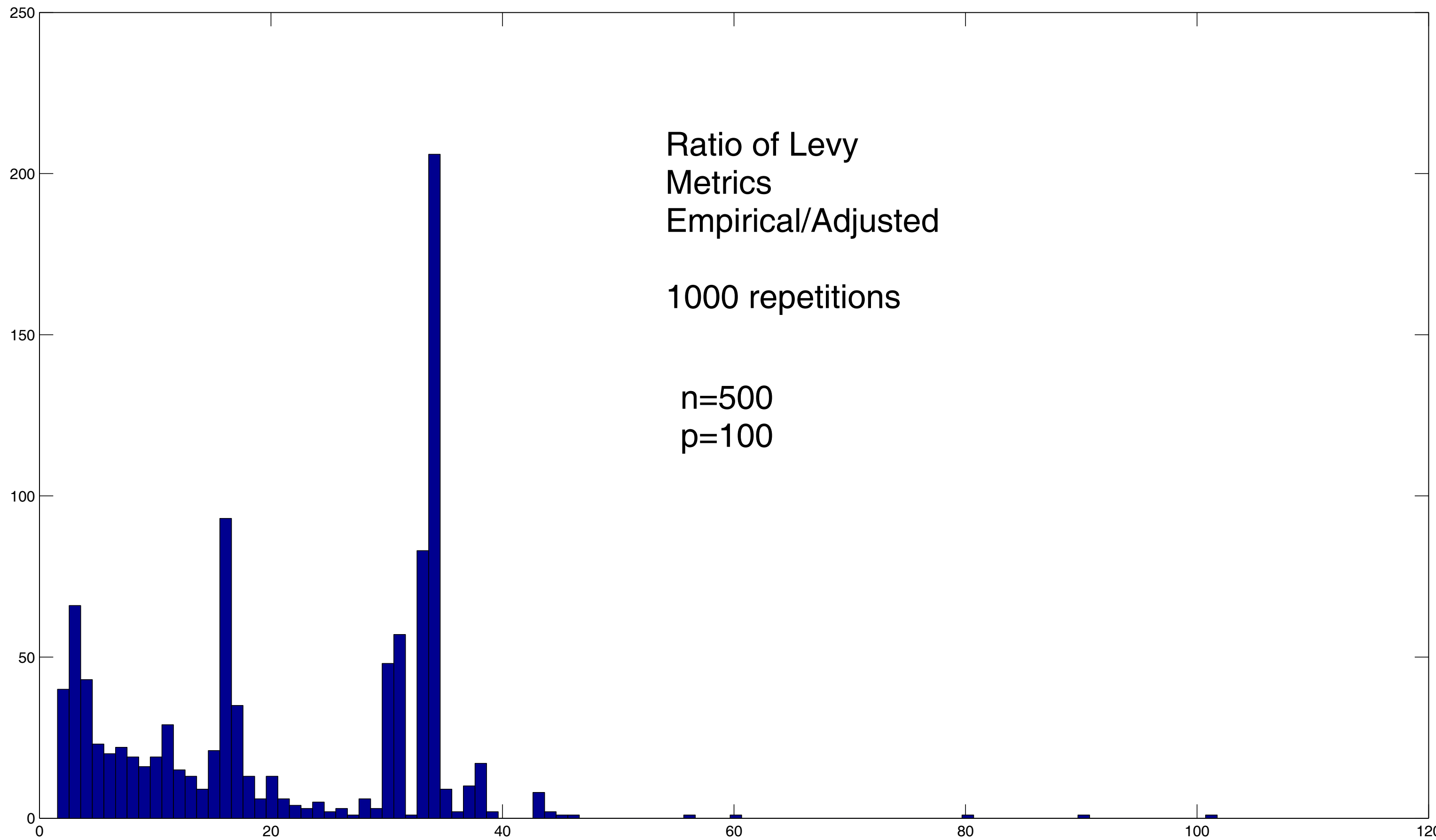


Figure 2: Ratio of Lévy metrics to Id

Example 2: population covariance = diagonal(50 1's, 50 2's) Population Spectrum = $1/2\delta_1 + 1/2\delta_2$

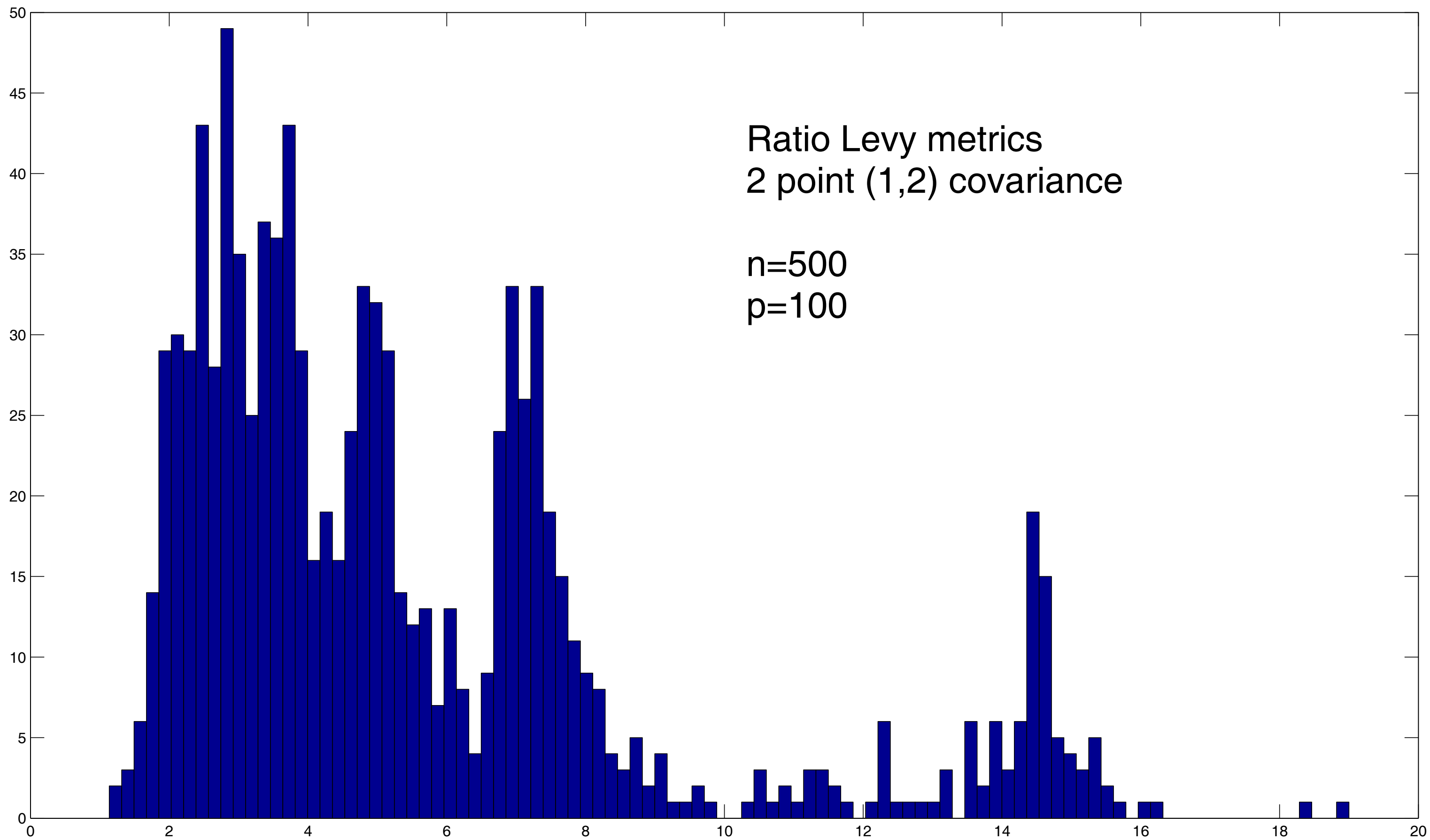


Figure 3: Ratio of Lévy metrics to population spectrum

Example 3: population covariance = Toeplitz($.3^k$), $k = 0$ to 100

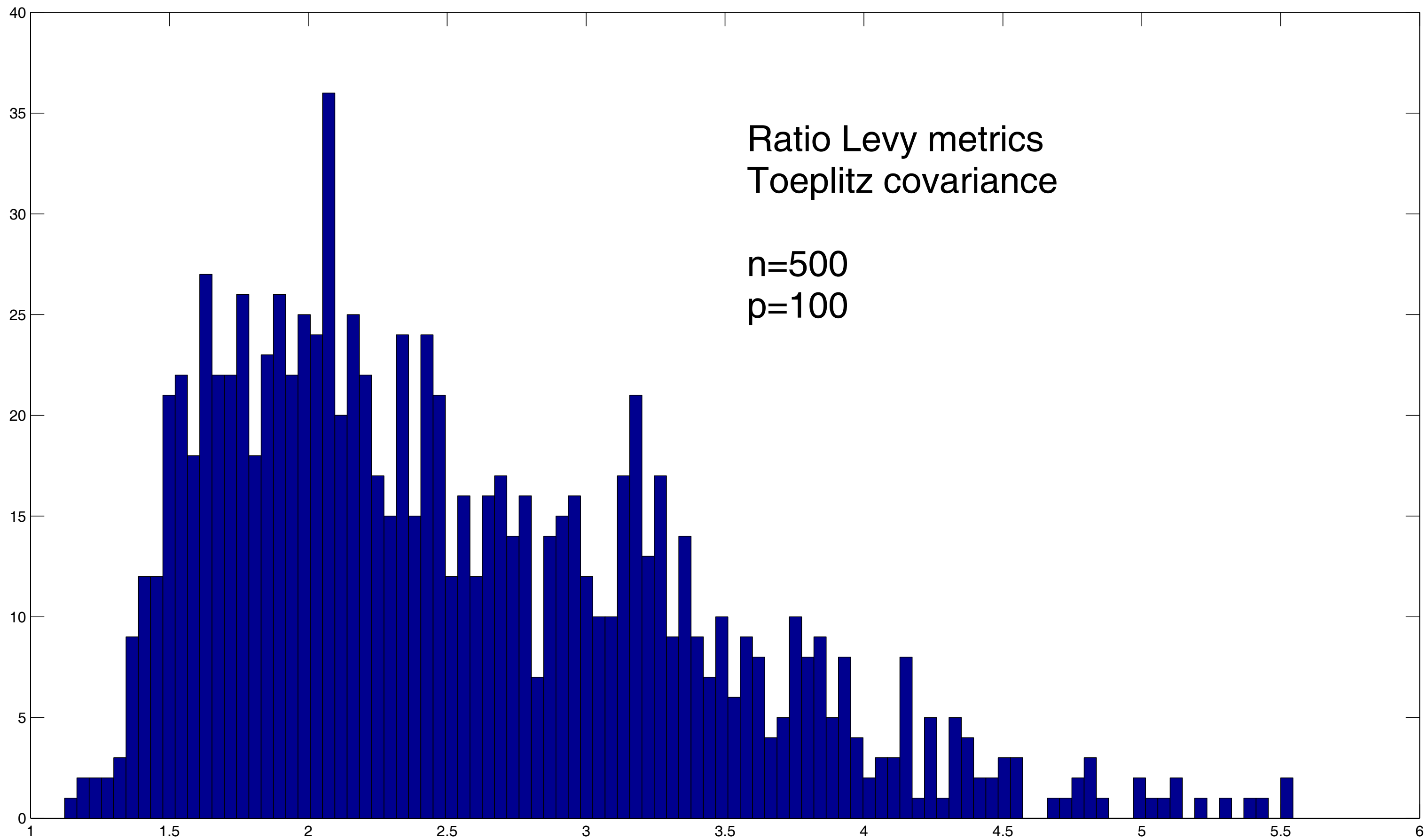


Figure 4: Ratio of Lévy metrics to population spectrum

Getting 2/3 rate: some elements of the proof

- Simplification:

  Lemma 1. [Seiler-Simon] *For trace class operators A and B, we have*

  $$|\det(\mathrm{Id} + A) - \det(\mathrm{Id} + B)| \leq \|A - B\|_1 \exp(1 + \|A\|_1 + \|B\|_1)$$

- Strategy is therefore to control $\|\tilde{S}_{n,p} - \bar{S}\|_1$, to first order.

Getting 2/3 rate: some elements of the proof

- Simplification:

  Lemma 1. [Seiler-Simon] *For trace class operators A and B, we have*

$$|\det(\mathrm{Id} + A) - \det(\mathrm{Id} + B)| \leq \|A - B\|_1 \exp(1 + \|A\|_1 + \|B\|_1)$$

- Strategy is therefore to control $\|\tilde{S}_{n,p} - \bar{S}\|_1$, to first order.
- $\tilde{S}_{n,p} = G_{n,p} H_{n,p} + H_{n,p} G_{n,p}$, $\bar{S} = 2G^2$

Getting 2/3 rate: some elements of the proof

- Simplification:

Lemma 1. [Seiler-Simon] *For trace class operators A and B, we have*

$$|\det(\mathrm{Id}+A)-\det(\mathrm{Id}+B)| \leq \|A-B\|_1 \exp(1+\|A\|_1+\|B\|_1)$$

- Strategy is therefore to control $\|\tilde{S}_{n,p}-\bar{S}\|_1$, to first order.

- $\tilde{S}_{n,p}=G_{n,p}H_{n,p}+H_{n,p}G_{n,p}$, $\bar{S}=2G^2$

- Getting $2/3$ rate : what matters is $G_{n,p}+H_{n,p}$.

- Involved centering and scaling come from this trade-off

More details about the rate

- Hilbert-Schmidt norm easier to deal with than trace class norm

- Try to make product of Hilbert-Schmidt operators appear

More details about the rate

- Hilbert-Schmidt norm easier to deal with than trace class norm
- Try to make product of Hilbert-Schmidt operators appear
- Simple remark: for operators A, B and G

$$AB + BA - 2G^2 =$$

$$\frac{1}{4}\left[(A+B-2G)(A+B+2G)+(A+B+2G)(A+B-2G)-2(A-B)^2\right]$$

More details about the rate

- Hilbert-Schmidt norm easier to deal with than trace class norm
- Try to make product of Hilbert-Schmidt operators appear
- Simple remark: for operators A, B and G

$$AB + BA - 2G^2 =$$

$$\frac{1}{4}\left[(A+B-2G)(A+B+2G) + (A+B+2G)(A+B-2G) - 2(A-B)^2\right]$$

Conclusion: if A, B and G are Hilbert-Schmidt

$$2\|AB + BA - 2G^2\|_1 \leq \|A+B-2G\|_2\|A+B+2G\|_2 + \|A-B\|_2^2$$