Random Matrix Theory:

Applications in High-Dimensional Statistics

Group D: Swapnaneel, Srijan, Sevantee

Indian Statistical Institute, Kolkata

December 2, 2024

Contents

1. Introduction
2. Theoretical Framework
3. Inference on Covariance Matrices
4. Application in PCA
5. Detecting Number of Signals
6. Application in Changepoint Detection
7. Conclusion and Further Directions

Introduction

Why RMT?

- Most of the traditional Multivariate statistical method assumes the data dimension p to be fixed or not too large compared to the number of data points.
- In the modern era most of the datasets from Genomics, Finance, Signal, and Image Processing violate that assumption.
- So it is challenging to apply those traditional statistical methods.
- Also many traditional methods become very demanding regarding necessary assumptions for the method to work nicely for high-dimensional datasets.
- In modern datasets often those assumptions are not satisfied.

Why RMT?

Consistency of the Sample PCs

Let $X_1, \cdots, X_n \overset{iid}{\sim} N_p(0, \Sigma)$ where Σ is positive definite with all distinct eigenvalues $\lambda_1 > \cdots \lambda_p > 0$. Let $l_{n,1} \geqslant \cdots \geqslant l_{n,p}$ be the eigenvalues of $\frac{1}{n}\sum_{i=1}^{n} X_i X_i^T$ and $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_p)$. Then

$$\sqrt{n}(l_n - \lambda) \xrightarrow{\mathcal{L}} N_p(0, 2\Lambda^2) \tag{1}$$

as $n \to \infty$ where $l_n = \begin{bmatrix} l_{n,1} \\ \vdots \\ l_{n,p} \end{bmatrix}$ and $\lambda = \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_p \end{bmatrix}$

- For example, when the eigenvalues of the population covariance matrix are distinct, the sample PCs are consistent.
- But the assumption of distinct eigenvalues is not often satisfied for verbal audio data used for speech recognition or noisy signals.

Why RMT - An Illustration through Spiked Covariance Model

- For those cases, the covariance matrix is in the form

$$\Sigma = \sum_{j=1}^{M} \lambda_j \theta_j \theta_j^* + \sigma^2 I_p$$

having M leading ***significant*** eigenvalues to be distinct and rest all the same: $(\lambda_1 + \sigma^2) > \cdots > (\lambda_M + \sigma^2) > \sigma^2 = \cdots = \sigma^2$

- For the above model if
  - $\gamma = \lim_{n \to \infty} \frac{p}{n} \in (0, \infty)$
  - an eigenvalue $\ell_j \leqslant 1 + \sqrt{\gamma}$ with arithmetic multiplicity 1

then the angle between then the angle between the $j-$th sample and population eigenvectors converges to $\frac{\pi}{2}$ almost surely.

- This essentially shows to which extent the sample principal components can be inconsistent for these models.

Why RMT

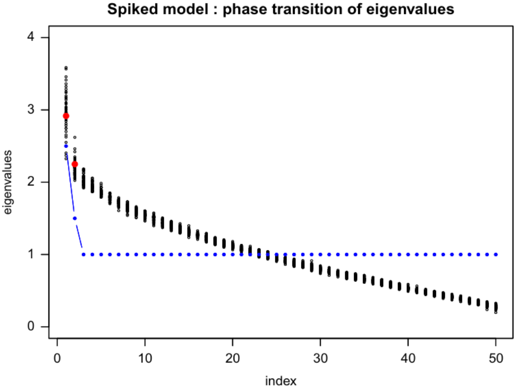


Figure: $p = 50$, $n = 200$ and eigenvalues of the covariance matrix are $\ell_1 = 2.5$, $\ell_2 = 1.5$, $\ell_j = 1$ for $j = 3, \ldots, p$. Blue dots correspond to the population eigenvalues. Black circles correspond to the sample eigenvalues (based on i.i.d. Gaussian samples) for 50 replicates. Solid red circles indicate the theoretical limits of the first two eigenvalues for $\gamma = p/n = 0.25$.

Why RMT?

RMT provides a unified framework to work on such problems.

Theoretical Framework

How to study Large Random Matrices

- In most cases, random matrices are studied in terms of their spectrum.
- For a random matrix $\mathbf{X}$ with eigenvalues $\lambda_1, \cdots, \lambda_n$ its **Empirical Spectral Distribution (ESD)** denotes the uniform distribution on $\lambda_1, \cdots, \lambda_n$.
- So for a random matrix $\mathbf{X}$ its **ESD** is a random measure taking value in $\mathbb{R}^*$ (set of all probability measures on $\mathbb{R}$).
- If $\mathbf{X}$ is real symmetric, one can write the empirical distribution function $F^{\mathbf{X}}(t)$ as

$$F^{\mathbf{X}}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(\lambda_j \leqslant t), \ \ t \in \mathbb{R}$$

- If $n \to \infty$, the limit of this distribution (if exists) is called **Limiting Spectral Distribution(LSD)**.
- For many special random matrices, the **LSD** turns out to be a deterministic measure !!!

Two famous Random Matrices - Wishart and $F-$type Matrix

- If $X_1, \cdots, X_n \overset{iid}{\sim} N_p(\mathbf{0}, \Sigma)$, and $\mathbf{X} = [X_1 : \cdots : X_n]$, then the distribution of $\mathbf{XX^T}$ is called Wishart distribution with parameter , degree of freedom n and dimension p and abbreviated as $W_p(\Sigma, m)$
- If $A \sim W_p(\Sigma, m)$ and $B \sim W_p(\Sigma, n)$ and A, B independent then the distribution of $B^{-1/2}AB^{-1/2}$ is called $F(p, m, n)$
- If the data $X_1, \cdots, X_n \overset{iid}{\sim} N_p(\mathbf{0}, \Sigma)$, and $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})^T$ is the sample covariance matrix, then $n\mathbf{S} \sim W_p(\Sigma, n-1)$.
- This makes these two distributions very frequently occur in many statistical problems such as inference on covariance matrix, PCA, MANOVA, High-dimensional Linear Models, High-dimensional Factor Models, Signal estimation from a noisy data, etc.

Study of the Spectrum

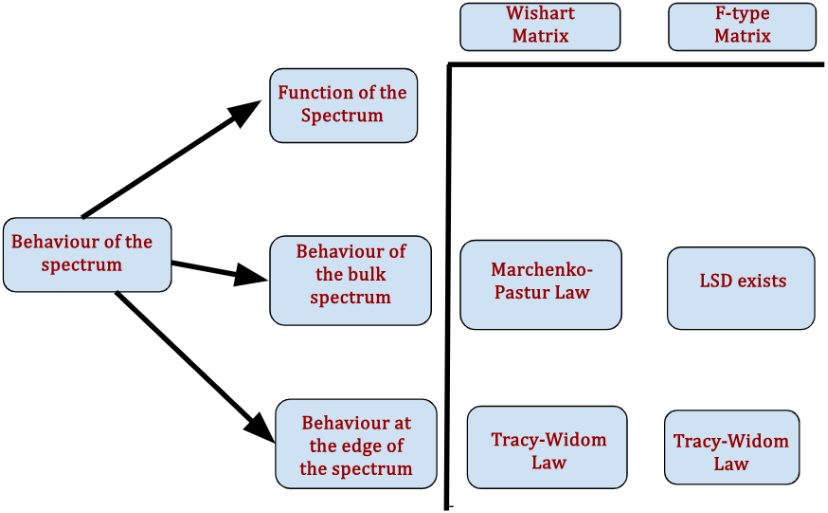


Figure: Asymptotic Properties of the Spectrum of Large Wishart and $F-$type matrices

Function of the Spectrum

Let $\mathbf{X}_n \sim W_p(\Sigma, n)$ where $n \geqslant p$. Let $\lambda_1^{(n)}, \cdots, \lambda_p^{(n)}$ be the eigenvalues of $\mathbf{X}_n$ and $\lambda_1, \cdots, \lambda_p$ be the eigenvalues of Σ. Then,

$$\sup_{x \in \mathbb{R}} \left| P\left(\sqrt{\frac{n}{2p}} \left(\sum_{i=1}^{p} \log\left(\frac{\lambda_i^{(n)}}{\lambda_i} \right) - \sum_{i=1}^{p} \log\left(n - p + i\right) \right) \leqslant x \right) - \Phi(x) \right| = O\left(\frac{p}{\sqrt{n}} \right)$$

where $\Phi(x)$ denotes the Standard Normal CDF.

So when $\frac{p}{\sqrt{n}} \to 0$, then

$$\sqrt{\frac{n}{2p}} \left(\sum_{i=1}^{p} \log\left(\frac{\lambda_i^{(n)}}{\lambda_i} \right) - \sum_{i=1}^{p} \log\left(n - p + i\right) \right) \xrightarrow{\mathcal{L}} N(0,1) \tag{2}$$

Behaviour of the bulk Spectrum for Wishart Matrices

- $\mathbf{X} \in \mathbb{R}^{p \times n}$ with i.i.d. real- or complex-valued entries with mean 0 and variance 1.
- $\lim\limits_{n \to \infty} \frac{p}{n} = \gamma \in (0, \infty)$
- Then as $n \to \infty$ as ESD of $\mathbf{S} = \frac{1}{n}\mathbf{X}\mathbf{X}^T$ converges to law F_γ.
- If $\gamma \in (0, 1]$, then F_γ has the p.d.f.:

$$f_\gamma(x) = \frac{\sqrt{(b_+(\gamma) - x)(x - b_-(\gamma))}}{2\pi\gamma x}, \quad b_-(\gamma) \leq x \leq b_+(\gamma), \tag{3}$$

  where

$$b_\pm(\gamma) = (1 \pm \sqrt{\gamma})^2 .$$

- For x outside this interval, $f_\gamma(x) = 0$.
- If $\gamma \in (1, \infty)$, then F_γ is a mixture of a point mass at 0 and the p.d.f. $f_{1/\gamma}(x)$, with weights $1 - 1/\gamma$ and $1/\gamma$, respectively.

Marcenko-Pastur Law

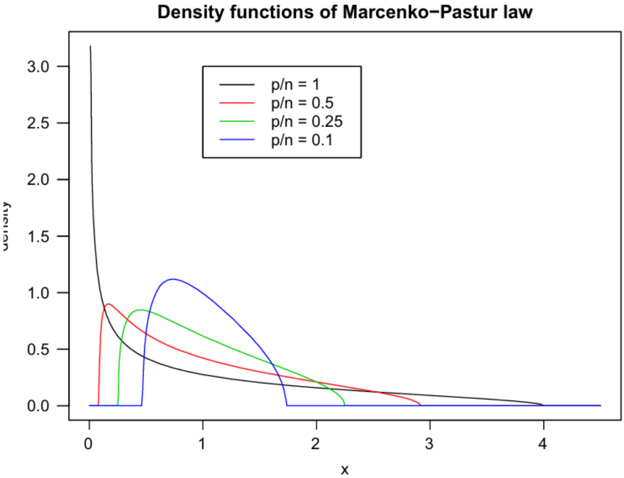
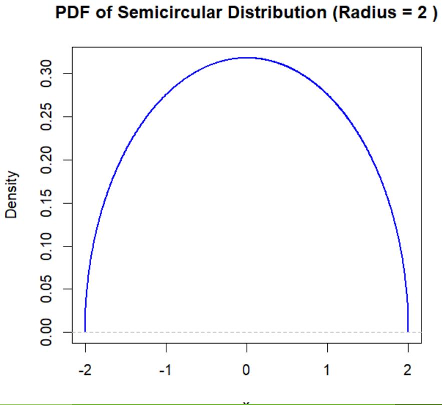


Figure: Marčenko–Pastur density functions for $\gamma = 0.1, 0.25, 0.5, 1$

Marcenko Pastur Law for $\gamma = 0$

- For $\gamma = 0$, the ESD of $\mathbf{S}_p = \frac{1}{2\sqrt{np}}\left(\mathbf{X}\mathbf{X}^T - nI_p\right)$ converges almost surely in distribution to semi-circular distribution, having the pdf

$$f(x) = \frac{1}{2\pi}\sqrt{4 - x^2}, \quad -2 \leq x \leq 2 \tag{4}$$



Behavior at the edge of Spectrum: Wishart Matrix

- $\mathbf{X}$ is a $p \times n$ matrix with i.i.d. real-valued entries with mean 0 and variance σ^2 and finite fourth moment.
- If $\gamma \in (0, \infty)$, $\lambda_{max}(\frac{1}{n}\mathbf{X}\mathbf{X}^T) \to (1+\sqrt{\gamma})^2\sigma^2$ a.s.
- If $\gamma \in (0, 1)$, $\lambda_{min}(\frac{1}{n}\mathbf{X}\mathbf{X}^T) \to (1-\sqrt{\gamma})^2\sigma^2$ a.s.

Tracy Widom Law

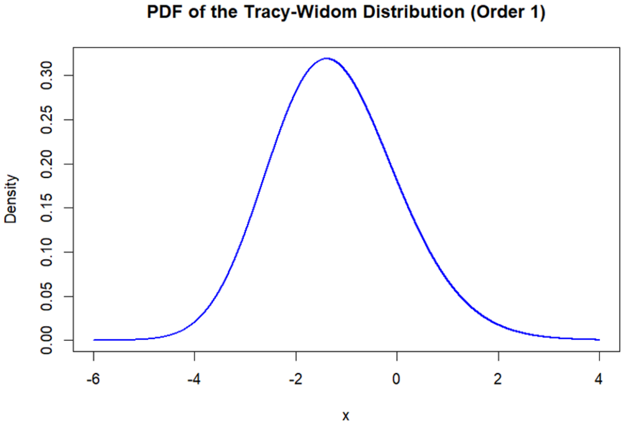
- If $\gamma \in (0, \infty)$, $\mathbf{X}_{i,j} \overset{i.i.d.}{\sim} N(0,1)$ then

$$\frac{l_1 - \mu_{n,p}}{\sigma_{n,p}} \xrightarrow{\mathcal{L}} W_1 \sim F_1$$

  F_1 CDF of Tracy-Widom Distribution.

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

Tracy Widom Distribution



Asymptotic properties of spectrum of $F-$ type matrices

Behaviour of Bulk Spectrum

- $\mathbf{F} = \mathbf{S}_m \mathbf{S}_n^{-1}$ with $\mathbf{S}_m$, $\mathbf{S}_n$ independent sample covariance matrix
- data dimension p, sample sizes m, n, data has mean 0, variance 1 and $\frac{p}{m} \to \gamma \in (0, \infty)$, $\frac{p}{n} \to \gamma' \in (0, 1)$
- Then LSD of $\mathbf{F}$ exists and has the pdf

$$f_{\gamma,\gamma'}(x) = \frac{(1-\gamma')\sqrt{(b-x)(x-a)}}{2\pi x(\gamma + x\gamma')}\mathbf{1}(a < x < b)$$

Behaviour at the edge of the spectrum

- Under the same set of assumptions, a normalized version of $\lambda_1(\mathbf{F})$ (highest eigenvalue) goes in distribution to Tracy-Widom Distribution.

Maximum eigenvalue of an $F-$type matrix

- $\gamma, \gamma' \in (0,1), \breve{m} = \max\{m, p\}, \quad \breve{n} = \min\{n, m+n-p\}, \quad \breve{p} = \min\{m, p\}$
- $$\sin^2\left(\frac{\gamma}{2}\right) = \frac{\min\{\breve{p}, \breve{n}\} - \frac{1}{2}}{\breve{m} + \breve{n} - 1}, \quad \sin^2\left(\frac{\psi}{2}\right) = \frac{\max\{\breve{p}, \breve{n}\} - \frac{1}{2}}{\breve{m} + \breve{n} - 1}$$
- $$\mu_{J,p} = \tan^2\left(\frac{\gamma + \psi}{2}\right)$$
- $$\sigma^3_{J,p} = \frac{16\mu^3_{J,p}}{(\breve{m} + \breve{n} - 1)^2} \cdot \frac{1}{\sin(\gamma)\sin(\psi)\sin^2(\gamma + \psi)}$$

Then,

$$\lim_{p \to \infty} P\left(\frac{\frac{\breve{n}}{\breve{m}}\lambda_1 - \mu_{J,p}}{\sigma_{J,p}} \leqslant s\right) = F_1(s)$$

$F(\cdot)$ CDF of Tracy-Widom Distribution.

Summary

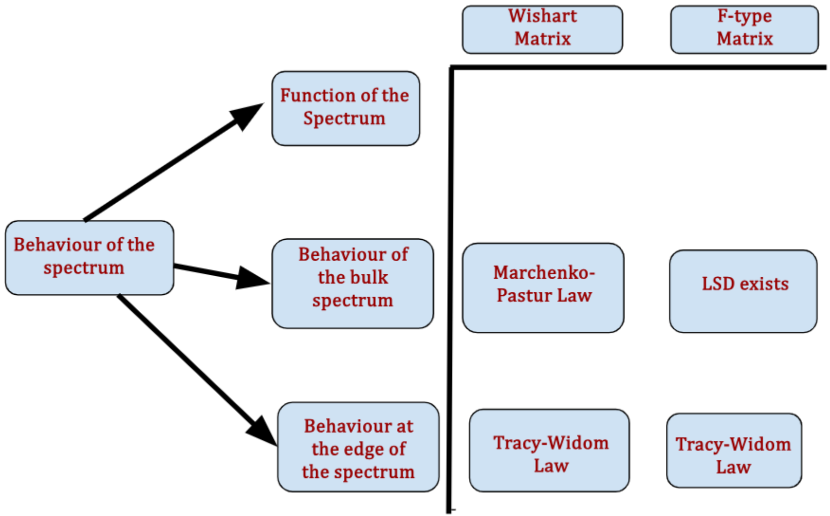


Figure: Asymptotic Properties of the Spectrum of Large Wishart and $F-$type matrices

Inference on Covariance Matrices

One Sample Testing for Covariance Matrix

- For $X_1, \cdots, X_n \overset{iid}{\sim} N_p(\mu, \Sigma)$ where μ, Σ are unknown.
- p, n are both large and $p \sim n^{\frac{1}{2}-\epsilon}$, $\epsilon > 0$
- Consider the testing problem with a two-sided alternative,

$$H_0 : \Sigma = \Sigma_0$$
$$H_1 : \Sigma \neq \Sigma_0$$

- $$\phi(\mathbf{X}) = \mathbf{1}\left(\sqrt{\frac{n-1}{2p}}\left|\sum_{i=1}^{p} \log\left(\frac{\hat{\lambda}_i}{\lambda_i}\right) - \sum_{i=1}^{p} \log(n-p+i)\right| > z_{\alpha/2}\right)$$

  is an asymptotic size α test.
- $\hat{\lambda}_1, \cdots, \hat{\lambda}_p$ are the eigenvalues of the sample covariance matrix.
- $\lambda_1, \cdots, \lambda_p$ are the eigenvalues of Σ_0

Test for High-dimensional Linear Model

- Consider the multivariate Linear Regression model,

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}$$

- $\mathbf{Y} = [Y_1 : \cdots : Y_m] \in \mathbb{R}^{n \times m}$ is the response matrix consisting n observations for each of the m response variable.
- $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the design matrix where p is the number of covariates.
- $\mathbf{E}$ denotes the error matrix, $\mathbf{E} \sim NDM(0, \Sigma)$ i.e. rows of $\mathbf{E}$ are iid from $N(0, \ )$.
- Consider the testing problem with a two-sided alternative,

$$H_0 : \Sigma = \Sigma_0$$
$$H_1 : \Sigma \neq \Sigma_0$$

Test for High-dimensional Linear Model

- Under H_0, SSE$=\mathbf{Y}^T(\mathbf{I}-\mathbf{P}_\mathbf{X})\mathbf{Y} \sim W_m(\Sigma_0, n-r)$, $r=\rho(\mathbf{X})$.
- An asymptotically level α test for H_0 is given by

$$\phi_{\mathcal{R}} := \mathbf{1}\left(\sqrt{\frac{n-r}{2m}}\left|\sum_{i=1}^{m}\log\left(\frac{\hat{\lambda}_i}{\lambda_i}\right) - \sum_{i=1}^{m}\log(n-r-k+i)\right| > z_{\alpha/2}\right)$$

- $\hat{\lambda}_1, \cdots, \hat{\lambda}_p$ are the eigenvalues of SSE $= \mathbf{Y}^T(\mathbf{I}-\mathbf{P}_\mathbf{X})\mathbf{Y}$
- $\lambda_1, \cdots, \lambda_p$ are the eigenvalues of Σ_0

Two Sample Test

- $X_1, \cdots, X_m \overset{iid}{\sim} N_p(\mu_1, \Sigma_1)$, $Y_1, \cdots, Y_n \overset{iid}{\sim} N_p(\mu_2, \Sigma_2)$, $\mu_1, \mu_2, \Sigma_1, \Sigma_2$ unknown
- $n \equiv n(m)$ satisfies $\lim_{m \to \infty} \frac{n}{m} = c \in (0, \infty)$.
- Consider the testing problem,

$$H_0 : \Sigma_1 = \Sigma_2$$
$$H_1 : \Sigma_1 \neq \Sigma_2$$

- An asymptotically level α test is,

$$\phi(\mathbf{X}, \mathbf{Y}) := \mathbf{1}\left(\sqrt{\frac{m}{2p\left(1+\frac{1}{c}\right)}}\left|\sum_{i=1}^{p} \log\left(\frac{\hat{\lambda}_i}{\hat{\lambda}_i^*}\right) - \sum_{i=1}^{p} \log\left(\frac{n-p+i}{m-p+i}\right)\right| > z_{\alpha/2}\right)$$

- $\hat{\lambda}_1, \cdots, \hat{\lambda}_p$ are the eigenvalues of $\mathbf{S_x}$
- $\hat{\lambda}_1^*, \cdots, \hat{\lambda}_p^*$ are the eigenvalues of $\mathbf{S_Y}$

Two Sample Test

- The same test can be done using the asymptotic theory of the largest root of $F-$type matrices as well.
- Consider,

$$\phi_{\mathcal{F}} := \mathbf{1}\left(\frac{\frac{\breve{n}}{\breve{m}}\lambda_1 - \mu_{J,p}}{\sigma_{J,p}} > F_1^{-1}(1-\alpha)\right)$$

  where $F_1(\cdot)$ CDF of Tracy-Widom Distribution.
- The center and scale parameters are a function of the sample sizes under H_0.
- λ_1 is the largest root of

$$\big((n-1)\mathbf{S_Y}\big)^{-1}\big((m-1)\mathbf{S_X}\big)$$

Wald Test for High Dimensional Regression

- Consider the multivariate Linear Regression model,

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

- $\mathbf{Y} \in \mathbb{R}^{n \times m}, \mathbf{X} \in \mathbb{R}^{n \times p}, \rho(\mathbf{X}) = p$,$\mathbf{E} \sim NDM(0, \Sigma)$
- Consider the wald's testing problem

$$H_0 : \mathbf{L^T B} = \mathbf{B_0}$$
$$H_1 : \mathbf{L^T B} \neq \mathbf{B_0}$$

- $\mathbf{L} \in \mathbb{R}^{p \times k}$, $\rho(\mathbf{L}) = k$.
- OLS estimate for $\mathbf{B}$ is given by $\widehat{\mathbf{B}} = (\mathbf{X^T X})^{-1} \mathbf{X^T Y}$

Wald Test for High Dimensional Regression

- under H_0,

$$\mathbf{A_p} := (\mathbf{L^T}\widehat{\mathbf{B}} - \mathbf{B_0})^{\mathbf{T}}(\mathbf{L^T}(\mathbf{X^T X})^{-\mathbf{1}}\mathbf{L})^{-\mathbf{1}}(\mathbf{L^T}\widehat{\mathbf{B}} - \mathbf{B_0}) \overset{H_o}{\sim} W_m(\Sigma, k)$$

- $\mathbf{B_p} = \mathbf{Y^T}(\mathbf{I} - \mathbf{P_X})\mathbf{Y} \sim W_m(\Sigma, n-p)$ and is independent with $\mathbf{A}_p$.
- Then using the maximum eigenvalue of $\mathbf{A}_p^{-1}\mathbf{B}_p$ one can find an asymptotic size α test.

Application in PCA

Spike Covariance Model

- $\Sigma = \sum_{j=1}^{M} \lambda_j \theta_j \theta_j^* + \sigma^2 I_p$
- $\theta_1, ..., \theta_M$ are orthonormal; $\lambda_1 \geq ... \geq \lambda_M > 0$ and $\sigma^2 > 0$.
- Σ has eigenvalues $\lambda_1 + \sigma^2, \cdots, \lambda_M + \sigma^2, \sigma^2, \cdots, \sigma^2$
- This model arises when the data has noise i.e. in verbal audio data, noisy signal, or in Factor Models.
- For the sake of simplicity σ^2 can be taken to be 1.

PCA For Spike Covariance Model

- $X_1, \cdots, X_n \overset{iid}{\sim} N_p(0, \Sigma)$ where Σ is a $p \times p$ positive definite matrix.
- Σ has eigenvalues $\ell_1 \geq \cdots \geq \ell_M > 1 = \cdots = 1$.
- $S = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T$ has eigenvalues $\hat{\ell}_1 \geq \cdots \geq \hat{\ell}_p$.
- $p, n \to \infty$ such that $\frac{p}{n} - \gamma = o(n^{-1/2})$ for a $\gamma \in (0, \infty)$.
- For a fixed $j \in \{1, 2, \cdots, M\}$ if $\ell_j > 1 + \sqrt{\gamma}$, then

$$\sqrt{n}\left(\hat{\ell}_j - \ell_j\left(1 + \frac{\gamma}{\ell_j - 1}\right)\right) \xrightarrow{\mathcal{L}} N(0, \sigma^2(\ell_j))$$

- $\sigma^2(\ell) := 2\ell^2\left(1 - \frac{\gamma}{(\ell-1)^2}\right)$

Detecting Number of Signals

The Setup

- We observe n i.i.d $p-$dimensional observations $\{\mathbf{x}_i\}_{i=1}^n$ from the model,

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \sigma\mathbf{n}(t)$$

  at the time points $t_1, \cdots, t_n$.

- $\mathbf{x}(t_1) = \begin{bmatrix} X_1(t_1) \\ \vdots \\ X_p(t_1) \end{bmatrix}, \cdots, \mathbf{x}(t_n) = \begin{bmatrix} X_1(t_n) \\ \vdots \\ X_p(t_n) \end{bmatrix}$
- $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_K] \in \mathbb{R}^{p \times K}$ have linearly independent columns.
- $s(t) = \begin{bmatrix} s_1(t) \\ \vdots \\ s_K(t) \end{bmatrix}$ is the vector of K source signals at time t.

Understanding the model

$$\boxed{\mathbf{x}(t)_{p\times 1} = \mathbf{A}_{p\times k}\mathbf{s}(t)_{k\times 1} + \sigma\mathbf{n}(t)_{p\times 1}}$$

- There K signal sources. At any timepoint t, the K signals are superimposed, get mixed with noise, and are sampled.
- For example consider audio data. K is usually NOT very large.

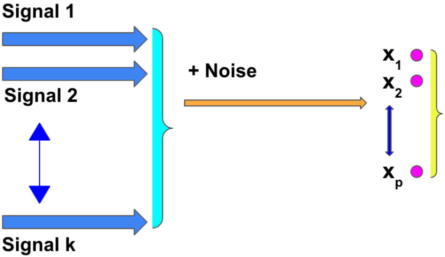


Figure: Observed at time points $t_1, t_2, \cdots, t_n$

Assumptions of the model

- Assume the signals are real-valued. The same set of results holds even if the signals are complex-valued as well.
- $s(t)$ is assumed to have $\boxed{\text{0 mean and stationary}}$ with $\boxed{\text{full rank covariance}}$ matrix.
- σ is the $\boxed{\text{unknown noise level}}$.
- $n(t)$ is a $p \times 1$ Gaussian noise vector, distributed $\boxed{N(0, I_p) \text{ and independent of } s(t)}$.

Under these assumptions, the population covariance matrix becomes

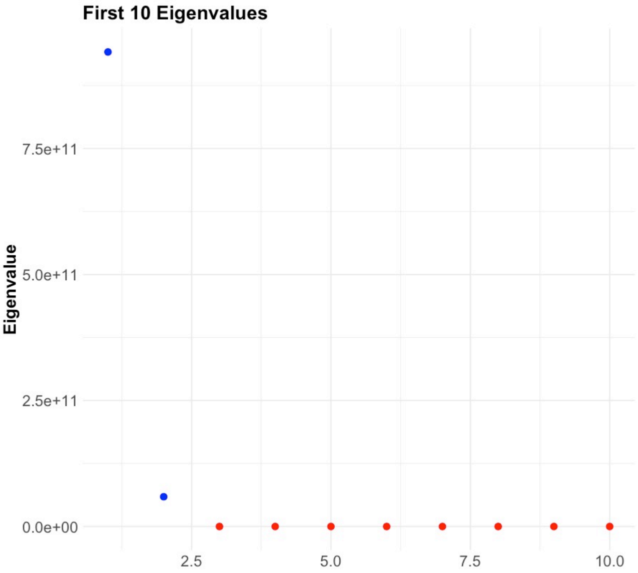
$$\boxed{\mathbf{W}^T \Sigma \mathbf{W} = \sigma^2 \mathbf{I}_p + \mathrm{diag}(\lambda_1, \ldots, \lambda_K, 0, \ldots, 0)}$$

Estimating the number of original signals

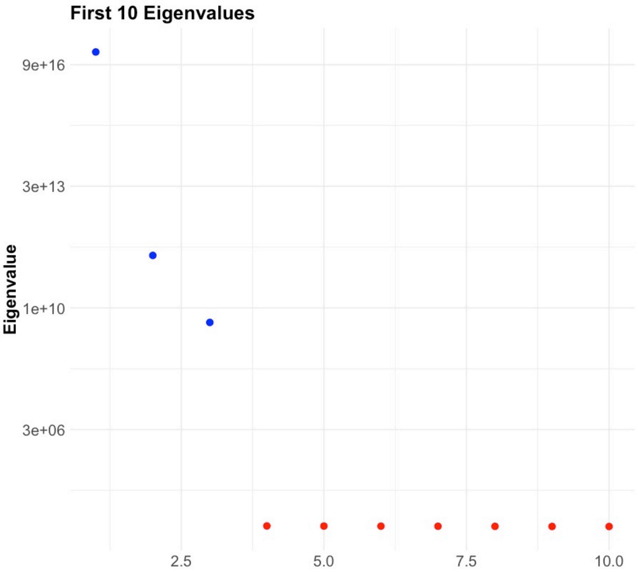
How can we determine the number of Signals from $\mathbf{x}(t_1), \cdots, \mathbf{x}(t_n)$?

- Eigenvalues of Σ are $\lambda_1 + \sigma^2, \cdots, \lambda_K + \sigma^2, \sigma^2, \cdots, \sigma^2$
- So the first K eigenvalues of Σ will be ***significantly*** large.
- We expect the same for $\mathbf{S}_n$.

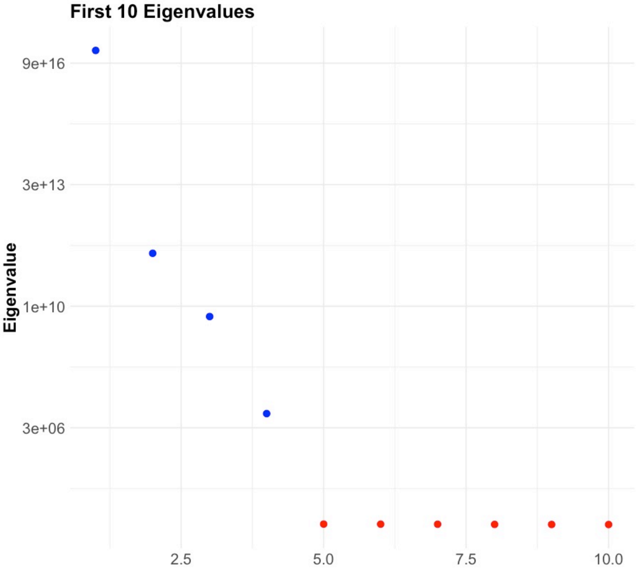
Eigenvalues for 2 signal case



Eigenvalues for 3 signal case



Eigenvalues for 4 signal case



Large ?

- So it turns out that the number of signals will essentially be the number of ***significantly*** large eigenvalues.
- How to determine the threshold of being large?

- Thanks to RMT!
- We have Tracy Widom Laws to determine asymptotic cut-off for the eigenvalues.
- So we will use this cut-off and only take the first few which pass the cut-off!!

More formally...

- We have eigenvalues $\boxed{l_1 \geq l_2 \geq \cdots \geq l_p \geq 0}$ of the sample covariance matrix

$$S_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}(t_i)\mathbf{x}(t_i)^T$$

- For $k = 1, 2, \cdots, min(p, n) - 1$, we test

$$\boxed{H_0\text{: at most } k-1 \text{ signals} \quad \text{vs} \quad H_1\text{: at least } k \text{ signals}}$$

- Under the null hypothesis, l_k arises from noise, so we

$$\boxed{\text{reject } H_0 \text{ if } l_k > \hat{\sigma}^2(k) C_{n,p,k}(\alpha)}$$

Fixing the threshold

- $\boxed{\hat{\sigma}^2(k) = \frac{1}{p-k}\sum_{j=k+1}^{p} l_j}$, estimate of the unknown noise level.
- $\boxed{C_{n,p,k}(\alpha) = \mu_{n,p-k} + s(\alpha)\xi_{n,p-k}}$
- $\boxed{\mu_{n,p} = \frac{1}{n}(\sqrt{n-1/2} + \sqrt{p-1/2})^2}$
- $\boxed{\xi_{n,p} = \sqrt{\frac{\mu_{n,p}}{n}}\left(\frac{1}{\sqrt{n-1/2}} + \frac{1}{\sqrt{p-1/2}}\right)^{1/3}}$
- $s(\alpha)$ is the $\boxed{1-\alpha \text{ quantile}}$ of the Tracy Widom distribution.

The Algorithm

- So, we will proceed sequentially for $k = 1, 2, \cdots, \min(p, n) - 1$
- Test the hypothesis at each stage.
- Will continue until the first non-rejection.

- Kritchman and Nadler (2009) showed

$$\Pr\{\text{reject } H_0 | H_0\} = \Pr\{\ell_k > \sigma^2 C_{n,p,k}(\alpha) | H_0\} \approx \alpha$$

- Also they showed that for a suitably chosen sequence of $\{\alpha\}_n$, $\hat{K}_{RMT,n}$ can be shown to be consistent i.e.

$$\boxed{\lim_{n \to \infty} \mathbb{P}(\hat{K}_{RMT,n} = K) = 1}$$

The Algorithm!

Algorithm 1: Algorithm for detecting number of signals

```
Input: Confidence level $\alpha$, observations $\ell_k$ for $k = 1, \ldots, \min(p, n) - 1$
Output: Estimated number of signals $\hat{K}_{\text{RMT}}$
for $k = 1$ to $\min(p, n) - 1$ do
    Compute the threshold $\hat{\sigma}^2(k) C_{n,p,k}(\alpha)$ if $\ell_k > \hat{\sigma}^2(k) C_{n,p,k}(\alpha)$ then
        conclude that there are at least $k$ signals and set $k = k + 1$ ;
    else
        conclude that there are at most $k - 1$ signals;
        Set $\hat{K}_{\text{RMT}} = k - 1$;
        break;
return $\hat{K}_{\text{RMT}} = \arg\min_k \left\{ \ell_k < \hat{\sigma}^2(k)(\mu_{n,p-k} + s(\alpha)\xi_{n,p-k}) \right\} - 1$;
```

Errors from $N(0, 1)$

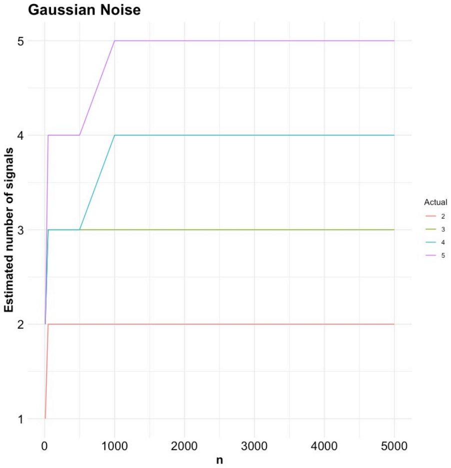


Figure: Actual vs Estimated number of Signals for varying n

Errors from t_5

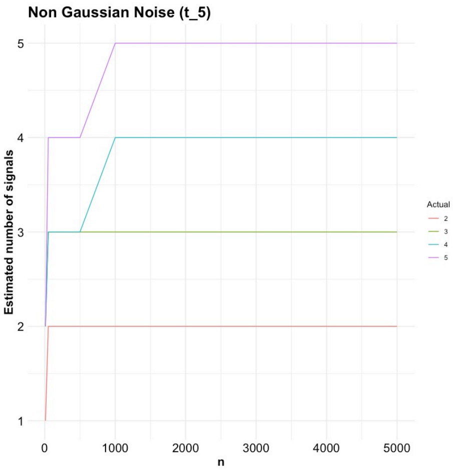


Figure: Actual vs Estimated number of Signals for varying n

Errors from $DoubleExp(1)$

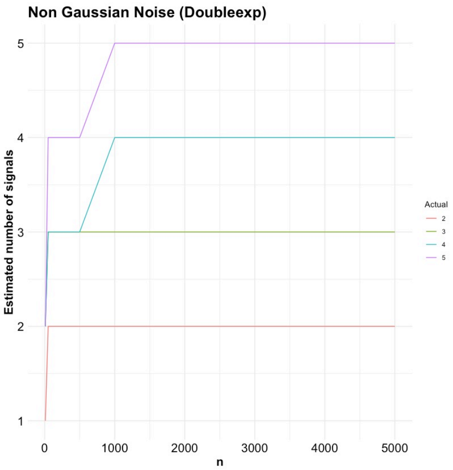


Figure: Actual vs Estimated number of Signals for varying n

Errors from $Cauchy(0, 1)$

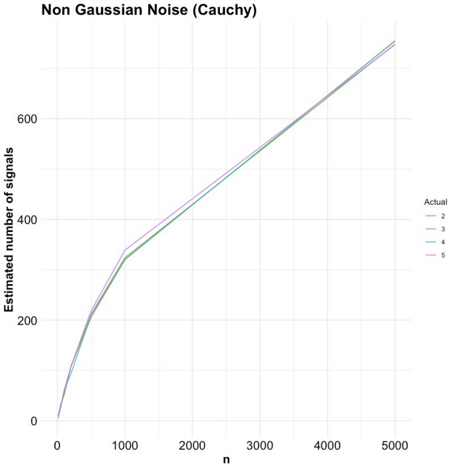


Figure: Actual vs Estimated number of Signals for varying n

Remarks

- For signals arising from the Cauchy distribution, the method tremendously overestimates the number of signals.
- Observe that the Tracy-Widom Law provides the threshold for significant eigenvalue.
- One of the required conditions for the largest eigenvalue to be asymptotically stable is the finiteness of the fourth moment.
- Here the eigenvalues correspond to the noise that even blows up for which a bunch of eigenvalues due to the noise also exceed the fixed threshold.
- For which, the method falsely classifies some noise variables as signals, so it overestimates the number of signals.

Application in Changepoint Detection

Observed data

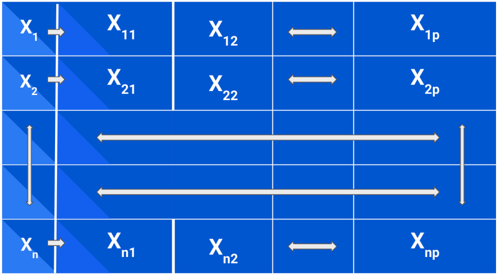
- Let $X_1, \dots, X_n \in \mathbb{R}^p$ be independent p-dimensional vectors with

$$\boxed{\mathrm{cov}(X_i) = \Sigma_{i,p}, \quad \text{for } 1 \le i \le n}$$

  where each $\Sigma_{i,p} \in \mathbb{R}^{p \times p}$ is of full rank.

- Furthermore, let $\mathbf{X}_{n,p}$ denote an $n \times p$ matrix defined by

$$\boxed{\mathbf{X}_{n,p} := (X_1^T, \dots, X_n^T)^T}$$



X_1 X_{11} X_{12} X_{1p}

X_2 X_{21} X_{22} X_{2p}

X_n X_{n1} X_{n2} X_{np}

Primary Goal

- Traditionally change point detection after observing the whole data is called "Offline" Change point detection.
- For now, let us consider the case of a single changepoint.
- So, for a known τ, we want to test

$$
\begin{aligned}
H_0 &: \Sigma_{1,p} = \cdots = \Sigma_{n,p} \\
H_1 &: \Sigma_{1,p} = \cdots = \Sigma_{\tau,p} \neq \Sigma_{\tau+1,p} = \ldots \Sigma_{n,p}
\end{aligned}
$$

An Example

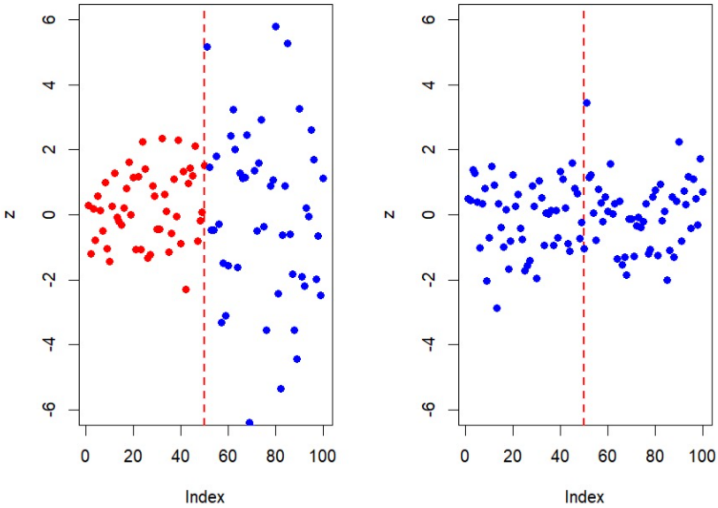


Figure: (Left) There is a changepoint at $\tau = 50$ and (Right) There is no change point

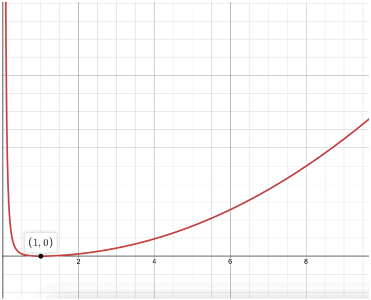
Thinking in terms of eigen values

How can we test whether a particular timepoint is a change point of covariances or not?

- Ryan and Killick (2023) constructs the two sample test statistics using the eigenvalues of the sample covariance matrices of the two samples.
- If matrices $\mathbf{A}$ and $\mathbf{B}$ are identical, all of the eigenvalues of $R(\mathbf{A}, \mathbf{B}) := \mathbf{B}^{-1}\mathbf{A}$ is 1.
- So, we can give a penalty each time when an eigen value goes far from 1 and define a distance between two matrices.

Let's look at the distance

$(1-x)^2 + \left(1 - \frac{1}{x}\right)^2$ has that property. So, we consider this function for each of the eigenvalues and consider their aggregate penalty.



So, we distance the following distance

$$\boxed{T(\mathbf{A},\mathbf{B}) = \sum_{j=1}^{p}[(1-\lambda_j(R(\mathbf{A},\mathbf{B})))^2 + \left(1-\lambda_j^{-1}(R(\mathbf{A},\mathbf{B}))\right)^2]}$$

Then?

- Nice! So we can take the sample estimates of the covariance matrices

$$\boxed{\frac{1}{n_1} X_{n_1,p}^T X_{n_1,p}, \frac{1}{n_2} X_{n_2,p}^T X_{n_2,p}}$$

  and take their distances.
- If that is large, we can say that the two population covariance matrices are significantly different.

Next?

- Ryan and Killick (2023) show that

$$\boxed{T\left(\frac{1}{n_1}X_{n_1,p}^T X_{n_1,p}, \frac{1}{n_2}X_{n_2,p}^T X_{n_2,p}\right) - p\int f^*(x)dF_\gamma(x) \to N(\mu(\gamma),\sigma^2(\gamma))}$$

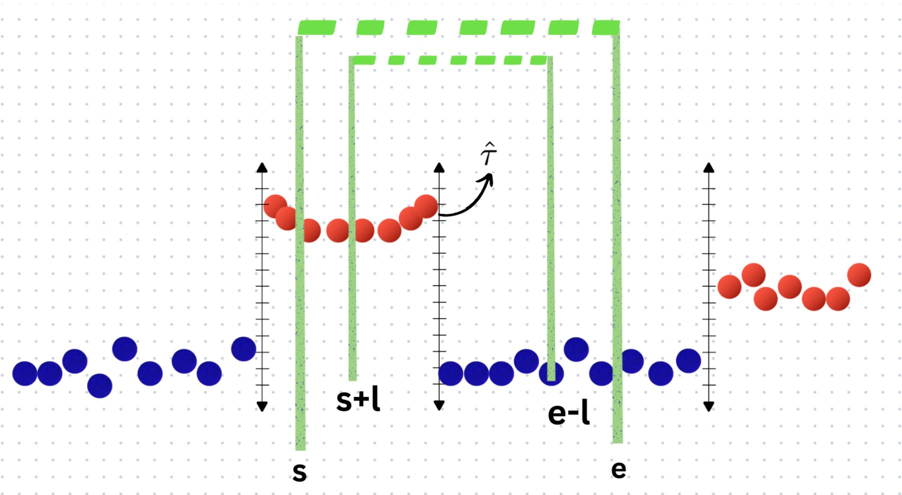
  for known functions $F_\gamma, f^*, \mu(\gamma), \sigma(\gamma), \gamma = (\gamma_1,\gamma_2), \frac{p}{n_1} \to \gamma_1, \frac{p}{n_2} \to \gamma_2$
- So, we can use the standard normal cutoff for testing after appropriately scaling and shifting.

Well! That solves one changepoint problem.
But how do we generalize for unknown multiple change-point problems?

Ratio Binary Segmentation, Ryan and Killick (2023)

- In an interval of time (s, e), for each timepoint τ in that range (expect edge deletion (l)) we find the normalized test statistic $\tilde{T}(\tau)$ by thresholding at τ.
- Then check whether the maximum one crosses the cutoff or not!
- If no change is found then the algorithm terminates.
- If a changepoint is found, it is added to the list of estimated changepoints, and the binary segmentation procedure is then run on the data to the left and right of the candidate change.
- This process continues until no more changes are found.

Visualizing the algorithm



Finally, The Algorithm!

Algorithm 2: Ratio Binary Segmentation (RatioBinSeg)

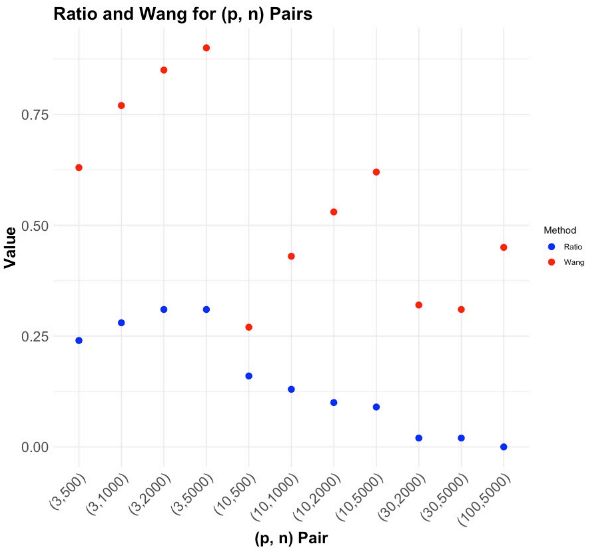
```
Input: Data matrix $X$, interval $(s, e)$, set of changepoints $C$, minimum
       segment length $\ell$, significance level $\alpha$
Output: Set of changepoints $C$
Set $\nu = 1 - \frac{\alpha}{n^2}$;
for $\tau = s + \ell$ to $e - \ell$ do
    Compute $\gamma := \left(\frac{p}{\tau}, \frac{p}{n-\tau}\right)$;
    Compute $\widetilde{T}(\tau) := \sigma^{-1/2}(\gamma)\left(T\left(\overline{\Sigma}(s,\tau), \overline{\Sigma}(\tau,e)\right) - p\int f^*(x)dF_y - \mu(\gamma)\right)$;
end
Set $\hat{\tau} := \arg\max_\tau \widetilde{T}(\tau)$ for $s + \ell < \tau < e - \ell$;
if $\widetilde{T}(\hat{\tau}) > \nu$ then
    Set $C_l :=$ RatioBinSeg$(X, (s, \hat{\tau}), C, \ell, \alpha)$;
    Set $C_r :=$ RatioBinSeg$(X, (\hat{\tau}, e), C, \ell, \alpha)$;
    Update $C = C \cup \{\hat{\tau}\} \cup C_l \cup C_r$;
end
return Set of changepoints $C$;
```

Simulation Study

- For the multiple changepoint setting, let $\tau := \{\tau_1, ..., \tau_m\}$ and $\hat{\tau} := \{\hat{\tau}_1, ..., \hat{\tau}_{\hat{m}}\}$ to denote the set of true changepoints and the set of estimated changepoints, respectively.
- The changepoint τ_i is said to be detected correctly if $|\hat{\tau}_j - \tau_i| \leq h$ for some $1 \leq j \leq \hat{m}$
- Denote the set of correctly estimated changes by $\hat{\tau}_c$.
- $h = 20$ is chosen for simulation, although it should be noted that in reality the desired accuracy would be application-specific and dependent on the minimum segment length l. Then the False Positive Rate (FPR) is defined as the number of wrongly detected changepoints out of the detected ones, i.e.

$$FPR = \frac{|\hat{\tau}| - |\hat{\tau}_c|}{|\hat{\tau}|}$$

Simulation Study



Conclusion and Further Directions

Conclusion

- We discussed the LSD and limits of extreme eigenvalues of Wishart and $F-$type matrices.
- We used the log-transformed eigenvalues for inference on covariance matrix in one-sample, two-sample, and high-dimensional regression setup.
- We discussed the behavior of Principal Components for Spike Covariance matrices.
- We used the Tracy-Widom laws to determine the number of signals from noisy data.
- A changepoint detection method for the covariance matrix was also demonstrated.

Further Directions

- Extensions of RMT theory for dependent data and apply in high-dimensional time series.
- Extensions of RMT theory for missing values and apply that in Spatiotemporal data when for each time point a couple of spatial observations are missing.
- Improve computational methods for large random matrix-based methods.

Thank you!

Reference

- Distribution of eigenvalues for some sets of random matrices: *Marchenko, Vladimir Alexandrovich and Pastur, Leonid Andreevich (1967)*
- Convergence to the semicircle law: *Bai, Z. D. and Yin, Y. Q. (1988)*
- On limit of the largest eigenvalue of the large dimensional sample covariance matrix: *Yin, Y., Bai, Z., and Krishnaiah, P. (1984)*
- Limit of the smallest eigenvalue of a large dimensional sample covariance matrix: *Bai, Z.-D. and Yin, Y.-Q. (2008)*
- On the distribution of the largest eigenvalue in principal components analysis: *Johnstone, I. M. (2001)*
- The tracy–widom law for the largest eigenvalue of f type matrices: *Han, X., Pan, G., and Zhang, B. (2016)*
- Multivariate analysis, volume 88. John Wiley & Sons: *Mardia, K. V., Kent, J. T., and Taylor, C. C. (2024)*
- Eigenvalues of large sample covariance matrices of spiked population models: *Baik, J. and Silverstein, J. W. (2006)*
- Asymptotics of sample eigenstructure for a large dimensional spiked covariance model: *Paul, D. (2007)*
- Detecting changes in covariance via random matrix theory: *Ryan, S. and Killick, R. (2023)*