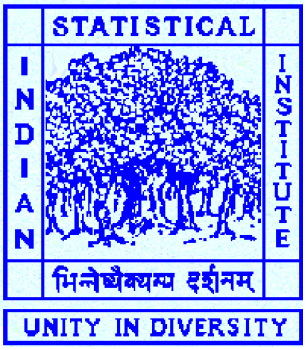
Indian Statistical Institute, Kolkata



Application of Random Matrix Theory in High-Dimensional Statistics

---

Swapnaneel Bhattacharyya (MB2427)
Srijan Chattopadhyay (MB2425)
Sevantee Basu (MB2421)
Group D

M.Stat First Semester

December 2, 2024

Application of Random Matrix Theory in High-Dimensional Statistics

Swapnaneel Bhattacharyya,[1,2], Srijan Chattopadhyay,[1,3], Sevantee Basu,[1,4]

[1] Indian Statistical Institute, 203 B.T. Road, Kolkata-700108, India

[2] swapnaneelbhattacharyya@gmail.com

[3] srijanchatterjee123456789@gmail.com

[4] sevanteebasu@gmail.com

Abstract: This article provides an overview of random matrix theory (RMT) with a focus on its growing impact on the formulation and inference of statistical models and methodologies. Emphasizing applications within high-dimensional statistics, we explore key theoretical results from RMT and their role in addressing challenges associated with high-dimensional data. The discussion highlights how advances in RMT have significantly influenced the development of statistical methods, particularly in areas such as covariance matrix inference, principal component analysis (PCA), signal processing, and changepoint detection, demonstrating the close interplay between theory and practice in modern high-dimensional statistical inference.

Key words: Random Matrix Theory(RMT), Empirical Spectral Distribution (ESD), Limiting Spectral Distribution (LSD), Principal Component Analysis (PCA), Canonical Correlation Analysis (CCA), Changepoint Detection (CPD)

Contents

1 Introduction

In recent years, the field of statistics has seen a significant shift driven by the rapid generation of large, complex datasets across diverse disciplines such as genomics, atmospheric science, communications, biomedical imaging, and economics. These datasets, often high-dimensional due to their representation in standard coordinate systems, pose challenges that extend beyond the scope of classical multivariate statistical methods. This evolving landscape has necessitated the integration of advanced mathematical frameworks, including convex analysis, Riemannian geometry, and combinatorics, into statistical methodologies. Among these, random matrix theory has emerged as a powerful tool for addressing key theoretical and practical problems in the analysis of high-dimensional data.

In this review article, we focus on several application areas of random matrix theory (RMT) in high-dimensional statistics. These include problems in dimension reduction, hypothesis testing for high-dimensional data, regression analysis, and covariance estimation. We also briefly describe the important role played by RMT in enabling certain theoretical analyses in wireless communications and changepoint detection. The challenges posed by high-dimensional data have sparked renewed interest in several classical phenomena within random matrix theory (RMT). Among these, the concept of universality holds particular significance, offering insights into the applicability of statistical techniques beyond the traditional framework based on the multivariate Gaussian distribution. This article focuses on aspects of RMT that are most relevant to statistical questions in this context. In particular, attention is directed toward the behavior of the bulk spectrum, represented by the empirical spectral distribution, and the edge of the spectrum, characterized by the extreme eigenvalues of random matrices. Given the central role of the sample covariance matrix in multivariate analysis, a significant portion of this work is devoted to examining its spectral properties and their implications for statistical applications.

A detailed discussion of these topics is present in Bai and Silverstein (2010), Couillet and Liao (2022). Also Johnstone (2006), Paul and Aue (2014) discuss many RMT-based approaches to modern high-dimensional problems. Though RMT has a wide variation of applications beyond statistics, e.g. wireless communications, finance, and econometrics, in this article our key focus is to discuss the RMT-based approaches to standard high-dimensional problems and their novelties in comparison to traditional methods. The article is organized as follows. In Section 3, we discuss the key theoretical results from random matrix theory which provides a framework for the statistical methods. We focus mainly on the asymptotic theory of the spectrum of the two kinds of random matrices - covariance matrix and the ratio of covariance matrices, for their widespread applications in statistics. For each of the two kinds of matrices, we discuss the properties of their bulk spectrum and behavior at the edge of the spectrum, when the matrices are of large dimensions. In the next Section 4, we discuss statistical applications of RMT. We focus on four problems: inference on covariance matrices, application in PCA, application in statistical signal detection removing noise, and changepoint detection.

Theorem 1 being our contribution, we present the proof of that theorem in the appendix section. We also present a few theoretical applications demonstrating the novelty of this result in Section 4.1. The rest of the theorems have appeared in the cited papers and we refer to those cited papers for their proof.

1.1 Notations and Abbreviations

In this paper, $\xrightarrow{\mathcal{P}}$ means convergence in probability, $\xrightarrow{\mathcal{L}}$ means convergence in distribution. For a random variable X, $F_X(\cdot)$ denotes its CDF. $\mathbf{1}(\cdot)$ denotes the indicator function. RMT means Random Matrix Theory, ESD stands for Empirical SPectral Distribution, LSD indicates Limiting Spectral Distribution. *as* μ means almost surely wrt the measure μ.

2 Background and Motivation

Random matrices play a fundamental role in statistical analysis, particularly in the study of multivariate data. Classical multivariate analysis, as detailed in influential works such as Mardia et al. (2024), and Muirhead (2009), frequently addresses key problems through the analysis of random matrices. These problems are typically formulated in terms of the eigen-decomposition of Hermitian or symmetric matrices and can be broadly classified into two categories.

The first category involves the eigen-analysis of a single Hermitian matrix, often referred to as the single Wishart problem, encompassing methods such as principal component analysis (PCA), factor analysis, and tests for population covariance matrices in one-sample problems. The second category includes generalized eigenvalue problems involving two independent Hermitian matrices of the same dimension, commonly known as the double Wishart problem. This includes applications like multivariate analysis of variance (MANOVA), canonical correlation analysis (CCA), tests for equality of covariance matrices, and hypothesis testing in multivariate linear regression.

Beyond these, random matrices also play a natural role in defining and characterizing estimators in multivariate linear regression, classification (involving sample covariance matrices), and clustering (using pairwise distance or similarity matrices). The analysis of eigenvalues and eigenvectors of random symmetric or Hermitian matrices has a long history in statistics, dating back to Pearson (1901) pioneering work on dimensionality reduction through PCA. This article provides a concise overview of these classical problems to set the stage for the broader discussion of random matrix theory in statistical applications.

Principal component analysis (PCA) is a highly versatile nonparametric tool for data reduction and model building. The formulation of PCA in classical multivariate analysis at

the population level is as follows. Suppose that we measure p variables (assume real-valued, for simplicity), expressed as a random vector $\mathbf{X} = \left(X^{(1)}, \ldots, X^{(p)}\right)^T$. Suppose also that the random vector $\mathbf{X}$ has finite variance $\mathbf{\Sigma} = \mathbb{E}\big[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T\big]$. The primary goal of PCA is to obtain a lower-dimensional representation of the data in the form of linear transformations of the original variable, subject to the condition that the residual variance is as small as possible.

This can be achieved by considering a sequence of linear transformations given by $\mathbf{v}_k^T\mathbf{X}, k = 1, 2, \ldots, p$, satisfying the requirement that $\mathrm{Var}(\mathbf{v}_k^T\mathbf{X})$ is maximized subject to the conditions that $\mathbf{v}_k$ are unit norm vectors in $\mathbb{R}^p$ (or $\mathbb{C}^p$ if the data is complex-valued), and $\mathbf{v}_k$ is orthonormal to $\{\mathbf{v}_j : j = 1, \ldots, k-1\}$, i.e., $\mathbf{v}_k^T\mathbf{v}_j = 0$ for $j = 1, \ldots, k-1$. This optimization problem can be solved in terms of the spectral decomposition of the nonnegative definite Hermitian matrix $\mathbf{\Sigma}$:

$$\mathbf{\Sigma}\mathbf{v}_k = \ell_k\mathbf{v}_k, \quad k = 1, \ldots, p, \tag{2.1}$$

where $\mathbf{v}_k$ are the orthonormal vectors. Here, ℓ_k (always real-valued) is an eigenvalue associated with $\mathbf{v}_k$. Note that in this formulation the eigenvalues ℓ_k are ordered, i.e., $\ell_1 \geq \cdots \geq \ell_p \geq 0$. If ℓ_k is of multiplicity one, then $\mathbf{v}_k$ is unique up to a sign change.

In practice, we do not know $\mathbf{\Sigma}$ and we typically observe a sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ for the variable $\mathbf{X}$. In that case, the empirical version of PCA involves replacing $\mathbf{\Sigma}$ by its natural estimate $\mathbf{S}_n = (n-1)^{-1}\sum_{i=1}^n(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^T$, and performing the spectral decomposition for $\mathbf{S}_n$.

The corresponding eigenvectors $\widehat{\mathbf{v}}_k$ are often referred to as the sample principal components. The corresponding ordered eigenvalues $\widehat{\ell_k}$ are typically used to detect the dimension of the reduction subspace. One of the commonly used techniques is to plot the eigenvalues against their indices (so-called "scree plot") and then look for an "elbow" in the plot.

There are formal tests based on likelihood ratios (see, Mardia et al. (2024), Muirhead

(2009)) that assume that, after a certain index, the eigenvalues are all equal and that the observations are Gaussian. Notice that the name "single Wishart" arises from the fact that if $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d. $N_p(\mathbf{0}, \mathbf{\Sigma})$, then $(n-1)\mathbf{S}_n$ has $W_p(\mathbf{\Sigma}, n-1)$ distribution.

Under Gaussianity, one of the commonly used tests for sphericity, i.e., the hypothesis $H_0 : \mathbf{\Sigma} = \mathbf{I}_p$, is Roy's largest root test (Roy (1953)), which rejects H_0 if $\widehat{\ell}_1$, the largest eigenvalue of $\mathbf{S}$, exceeds a threshold determined by the level of significance. If $H_0 : \mathbf{\Sigma} = \sigma^2\mathbf{I}_p$ for some unknown σ^2, the corresponding generalized likelihood ratio test, under an alternative that assumes $\mathbf{\Sigma}$ to be a rank-one perturbation of $\sigma^2\mathbf{I}_p$, rejects for large values of $\widehat{\ell}_1/\big(\sum_{j=2}^{p}\widehat{\ell}_j\big)$ (Johnson and Graybill (1972); Nadler (2008)).

A *factor analysis problem* can be seen as a generalization of PCA in that it assumes a certain signal-plus-noise decomposition of the observation vector $\mathbf{X}$:

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{L}\mathbf{f} + \boldsymbol{\varepsilon}, \tag{2.2}$$

where $\mathbf{f}$ is an $m \times 1$ dimensional random vector, $\mathbf{L}$ is a $p \times m$ dimensional nonrandom matrix, $\mathbf{f}$ and $\boldsymbol{\varepsilon}$ are uncorrelated, and $\boldsymbol{\varepsilon}$ has mean 0 and variance $\mathbf{\Psi}$, a $p \times p$ diagonal matrix. For identifiability, it is typically assumed that $\mathbb{E}[\mathbf{f}] = 0$ and $\mathbb{E}[\mathbf{f}\mathbf{f}^T] = \mathbf{I}_m$.

Under this setting, the covariance matrix of $\mathbf{X}$ is of the form $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^T + \mathbf{\Psi}$. Thus, if $\mathbf{\Psi}$ is a multiple of the identity, the problem of estimating $\mathbf{L}$ from data can be formulated in terms of a PCA of the sample covariance matrix. One important distinction between PCA and factor analysis is that, in the latter case, the practitioner implicitly assumes a causal model for the data. In general, factor analysis problems are often solved through a maximum likelihood approach (see Tipping and Bishop (1999)). A more enhanced version of the factor analysis model, the so-called dynamic factor model, is used extensively in econometrics, where the factors $\mathbf{f}$ are taken to be time-dependent (Forni et al. (2000)).

A detailed discussion of various versions of the double Wishart eigenproblem, including

a summary of the associated distribution theory when the observations are Gaussian, can be found in Johnstone and Lu (2009). Within this framework, we first consider the canonical correlation analysis (CCA) problem. Again, first, we deal with the formulation at the population level. Suppose that real-valued random vectors $\mathbf{X}$ and $\mathbf{Y}$ are jointly observed, where $\mathbf{X}$ is of dimension p and $\mathbf{Y}$ is of dimension q. Then a generalization of the notion of correlation between $\mathbf{X}$ and $\mathbf{Y}$ is expressed in terms of the sequence of canonical correlation coefficients defined as

$$\rho_k = \max_{(\mathbf{u},\mathbf{v})\in S_k} |\mathrm{Cor}(\mathbf{u}^\top \mathbf{X}, \mathbf{v}^\top \mathbf{Y})|, \quad k = 1, 2, \ldots, \min\{p, q\}, \tag{2.3}$$

where

$$S_k := \{(\mathbf{u}, \mathbf{v}) \in \mathbb{R}^{p+q} : \mathbf{u}^\top \mathbf{\Sigma_{XX}} \mathbf{u} = \mathbf{v}^\top \mathbf{\Sigma_{YY}} \mathbf{v} = 1;\ \mathbf{u}^\top \mathbf{\Sigma_{XX}} \mathbf{u}_j = \mathbf{v}^\top \mathbf{\Sigma_{YY}} \mathbf{v}_j = 0,\ j = 1, \ldots, k-1\},$$

with $\mathbf{\Sigma_{XX}} = \mathrm{Var}(\mathbf{X})$, $\mathbf{\Sigma_{YY}} = \mathrm{Var}(\mathbf{Y})$, and $(\mathbf{u}_k, \mathbf{v}_k)$ denoting the pair of vectors for which the maximum in (2.3) is attained. If $\mathbf{\Sigma_{XY}} = \mathrm{Cov}(\mathbf{X}, \mathbf{Y})$, then the optimization problem (2.3) can be formulated as the following generalized eigenvalue problem: the successive canonical correlations $\rho_1 \geq \cdots \geq \rho_{\min\{p,q\}} \geq 0$ satisfy the generalized eigen-equations

$$\det(\mathbf{\Sigma_{XY}} \mathbf{\Sigma_{YY}}^{-1} \mathbf{\Sigma_{YX}} - \rho^2 \mathbf{\Sigma_{XX}}) = 0. \tag{2.4}$$

When we have n samples $\{(\mathbf{X}_i, \mathbf{Y}_i) : i = 1, \ldots, n\}$ we can replace $\mathbf{\Sigma_{XX}}, \mathbf{\Sigma_{XY}}$ and $\mathbf{\Sigma_{YY}}$ by their sample counterparts and, assuming $n > \max\{p, q\}$, the corresponding sample canonical correlations $r_1 \geq \cdots \geq r_{\min\{p,q\}} \geq 0$ satisfy the sample version of Equation (2.4). It is shown in Mardia et al. (2024) that in the latter case, we can reformulate the corresponding generalized eigenanalysis problem as solving

$$\det(\mathbf{U} - r^2(\mathbf{U} + \mathbf{V})) = 0, \tag{2.5}$$

where $\mathbf{U}$ and $\mathbf{V}$ are independent Wishart matrices if $(\mathbf{X}_i, \mathbf{Y}_i)$ are i.i.d. Gaussian and $\mathbf{\Sigma_{XY}} = \mathbf{0}$, i.e., $\mathbf{X}$ and $\mathbf{Y}$ are independently distributed.

Next, we consider the multivariate Linear Regression model,

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{2.6}$$

where $\mathbf{Y} = [Y_1 : \cdots : Y_m] \in \mathbb{R}^{n\times m}$ 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. For inference purposes, it is further assumed that $\mathbf{E} \sim NDM(0, \mathbf{\Sigma})$ i.e. rows of $\mathbf{E}$ are iid from $N(0, \mathbf{\Sigma})$. Then, as described in Mardia et al. (2024), the union-intersection test for the linear hypothesis of the form $H_0 : \mathbf{CBD} = \mathbf{0}$ where $\mathbf{C}$ and $\mathbf{D}$ are specified conformable matrices, can be expressed in terms of the largest eigenvalue of $\mathbf{U}(\mathbf{U}+\mathbf{V})^{-1}$ where $\mathbf{U}$ and $\mathbf{V}$ are appropriately specified independent Wishart matrices (under Gaussianity of the entries of $\mathbf{E}$).

The two-sample test for equality of variances assumes that we have i.i.d. samples from two normal populations $N_p(\boldsymbol{\mu}_1, \mathbf{\Sigma}_1)$ and $N_p(\boldsymbol{\mu}_2, \mathbf{\Sigma}_2)$ of sizes n_1 and n_2, say. Then several tests for the hypothesis $H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ can be formulated in terms of functionals of the eigenvalues of $\mathbf{U}(\mathbf{U}+\mathbf{V})^{-1}$ where $\mathbf{U} = (n_1 - 1)\mathbf{S}_1$ and $\mathbf{V} = (n_2 - 1)\mathbf{S}_2$ are the sample covariances for the two samples, which would follow independent Wishart distributions in p dimensions with d.f. $n_1 - 1$ and $n_2 - 1$ and dispersion matrix $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ under H_0.

Now it is to be noted that the traditional methods to deal with the above problems assume the dimension of the data to be fixed and relatively small compared to the number of data points. But in the modern era, most of the high-dimensional data arising in fields such as genomics, economics, atmospheric science, chemometrics, and astronomy, to name a few, are of enormously large dimensions which makes it very challenging to apply the traditional methods directly to those datasets. And so, to accommodate the analysis of such datasets, it is imperative to either modify or reformulate some of the statistical techniques. This is where RMT has been playing a significant role, especially over the last decade. In the next sections, we develop these modern RMT-based methods with a key focus on their

applications in statistics.

3 High Dimensional Random Matrices

In Random Matrix Theory, two particular kinds of random matrices draw special attention for their remarkable application in statistics - Covariance Matrices and F-type Matrices. In this section, we discuss the theoretical properties of these two kinds of random matrices, which play a key role in most modern developments in high-dimensional statistics. Most of these results focus on the spectrum's behavior when the matrix has a large dimension. Hence to address those properties, we first give a basic introduction to these matrix models and describe a couple of key questions associated with it.

In classical random matrix theory, in the context of covariance matrices, one of the most fundamental and crucially studied matrices is the Wishart Matrix. The Wishart matrix is defined as specifying two sequences of integers n, the sample size, and $p = p(n)$, the data dimension. Most of the results additionally assume that the sequences are related so that, as $n \to \infty$, $p = p(n) \to \infty$ satisfying,

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

So if, $X_1, \cdots, X_n \overset{iid}{\sim} N_p(\mathbf{0}, \mathbf{\Sigma})$, and $\mathbf{X} = [X_1 : \cdots : X_n]$, then the distribution of $\mathbf{X}\mathbf{X}^\mathbf{T}$ is called Wishart distribution with parameter $\mathbf{\Sigma}$, degree of freedom n and dimension p and abbreviated as $W_p(\mathbf{\Sigma}, m)$. A density of the distribution is given by

$$f_W(\mathbf{X}) = \frac{|\mathbf{X}|^{(n-p-1)/2} e^{-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}\mathbf{X})}}{2^{np/2}|\Sigma|^{n/2}\Gamma_p\left(\frac{n}{2}\right)}$$

where $\Gamma_p(\cdot)$ is the multivariate gamma function. The distribution was first studied by Wishart (1928) and continues to be a fundamental focus in multivariate statistics thereafter. The central reason for Wishart matrices being so frequent and useful in statistics is

their association with sample covariance matrices - the sample covariance matrix of a normal random sample follows a Wishart distribution i.e. If the data $X_1, \cdots, X_n \overset{iid}{\sim} N_p(\mathbf{0}, \mathbf{\Sigma})$, 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 $n\mathbf{S} \sim W_p(\mathbf{\Sigma}, n-1)$. Further detailed properties of the distribution can be found in Muirhead (2009), Mardia et al. (2024). From the definition, it can be seen that the Wishart distribution is the analog of the chi-square distribution of the univariate case. Similarly, the univariate $F-$distribution has also an extension for matrices, which is called Matrix $F-$distribution. If $\mathbf{A} \sim W_p(I_p, \nu)$ and $\mathbf{B} \sim W_p(I_p, \delta)$ and $\mathbf{A}, \mathbf{B}$ are independent then the distribution of $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$ is called a matrix $\mathbf{F}(I_p, \nu, \delta)$ distribution. The distribution was originally derived by Olkin and Rubin (1964). Perlman (1977) discusses several interesting properties of this distribution. Usually if $\mathbf{A}, \mathbf{B}$ are independent random matrices such that $\mathbf{A} \sim W_p(\mathbf{\Sigma}, m)$ and $\mathbf{B} \sim W_p(\mathbf{\Sigma}, n)$ and Σ is positive definite and $m \geqslant p$, then $\mathbf{A}^{-1}\mathbf{B}$ is called an $F-$type matrix in the literature. For their widespread applications such as in Linear Discriminant Analysis, Canonical Correlation Analysis, etc, $F-$type matrices are also extensively studied. However, the properties of these matrices, which played a crucial role in statistical inference and related fields over decades, have been studied beyond the parametric framework, under minimalistic assumptions mostly for the high-dimensional setup. We discuss the properties of those matrices in terms of their spectrum in both parametric and nonparametric frameworks.

3.1 Spectral Properties of large Sample Covariance Matrices

The sample covariance matrix is one of the most important random matrices in multivariate statistical inference. It is fundamental in hypothesis testing, principal component analysis, factor analysis, and discrimination analysis. Many test statistics are defined by their eigenvalues. However, for large matrices, it is more convenient to study the asymptotic behavior of their spectrum as they exhibit nice properties.

3.1.1 Properties of the whole Spectrum

Suppose $\mathbf{X}$ is a $n \times n$ random matrix having eignevalues $\lambda_1, \cdots, \lambda_n \in \mathbb{C}$. Then the empirical distribution of the eigenvalues of $\mathbf{X}$ is called the empirical spectral distribution(ESD) of $\mathbf{X}$. If the matrix $\mathbf{X}$ is real, symmetric then all of its eigenvalues are real and hence the empirical spectral distribution is given by $\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(\lambda_i \leqslant x)$ which is the case for Wishart matrices. In random matrix theory, the ESD is crucial to study as most of the properties of a matrix can be reformulated in terms of its eigenvalues and hence is of key interest. In different domains like machine learning, signal processing, and wireless communications, several functions of the eigenvalues give important objects to study (eg. Tulino et al. (2004), Couillet and Liao (2022) etc).

In the parametric setup, under the normality assumption of the data, when we have the exact distribution of the sample covariance matrix, one of the most fundamental questions one can ask is how to characterize the joint distribution of the spectrum. If $\mathbf{X} \sim W_p(\mathbf{\Sigma}, n)$, then the rank of $\mathbf{X}$ is $\min\{p, n\}$ almost surely. Now for $n \leqslant p-1$, the rank of $\mathbf{X} \leqslant p-1$ and hence the eigenvalues do not have a joint pdf. However, for $n > p-1$, the joint probability density function (pdf) for the eigenvalues of $\mathbf{X}$ exists and can be found in Muirhead (2009) and a detailed study on their joint distribution can be found in James and Lee (2014). Furthermore, for $n > p-1$, the following central limit theorem holds for log-transformed eigenvalues of $\mathbf{X}$.

Theorem 1. *Let $\mathbf{X}_n \sim W_p(\mathbf{\Sigma}, n)$ 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 $\mathbf{\Sigma}$. 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.

The proof of Theorem 1 can be found in the appendix. Hence, for $\frac{p}{\sqrt{n}} \to 0$ as $n \to \infty$,

we have

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

It is to be noted that the above central limit theorem is very useful for one-sample and two-sample testing for covariance matrices, to approximate the power function of such tests, and also for inference on covariance matrices in high dimensional linear models. A detailed discussion of these applications can be found in Section 4.1. The theorem also provides a rate of convergence of the eigenvalues of the sample covariance matrix to the population covariance matrix. Often for approximation purposes, functions of the spectrum of the population covariance matrices are estimated using that of the corresponding sample covariance matrix. In such cases, the theorem provides an upper bound of the error. Hence, it is useful for sample size determination if there is a predetermined allowable upper bound on the error.

In this context, a natural follow-up question is whether the weak limit of the ESD for Wishart matrices exists. The celebrated *Marcenko-Pastur Law* answers this question in the context of sample covariance matrices. With an assumption of the finiteness of the fourth moment of the entries of the data matrix, Marchenko and Pastur (1967) showed that depending on the value of $\gamma = \lim_{n\to\infty}\frac{p}{n}$, the weak limit of the ESD of sample covariance matrices exist.

Theorem 2 (Marcenko-Pastur Law)**.** *Suppose that* $\mathbf{X}$ *is a* $p \times n$ *matrix with i.i.d. real- or complex-valued entries with mean 0 and variance 1. Suppose* $\lim_{n\to\infty}\frac{p}{n} = \gamma \in (0,\infty)$ *. Then, as* $n \to \infty$*, the empirical spectral distribution (ESD) of* $\mathbf{S} = \frac{1}{n}\mathbf{X}\mathbf{X}^T$ *converges almost surely in distribution to a nonrandom distribution, known as the Marcenko–Pastur law and denoted by* F_γ*. If* $\gamma \in (0,1]$*, then* F_γ *has the p.d.f.:*

$$f_\gamma(x) = \frac{\sqrt{(b_+(\gamma)-x)(x-b_-(\gamma))}}{2\pi\gamma x}, \quad b_-(\gamma) \le x \le b_+(\gamma), \tag{3.1}$$

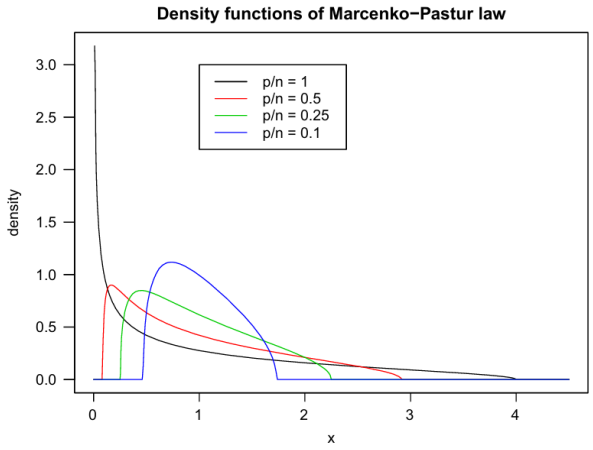


Figure 1: Marcenko–Pastur density functions for $\gamma = 0.1, 0.25, 0.5, 1$

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.

It is to be noted that the above result is distribution-free, in the sense the limiting distribution only depends on the limiting ratio of data dimension and sample size (γ) and is free of the data distribution. As γ increases from 0 to 1, the spread of the eigenvalues also increases. However, a necessary condition for the weak limit of the ESD to exist is to $\gamma > 0$. For $\gamma = 0$, as illustrated in Figure 1, the maximum and minimum eigenvalues converge to 1 and hence Marcenko-Pastur law does not hold for this case. However, with an assumption of the finiteness of the fourth moment of the entries of $\mathbf{X}$, applying a suitable centering and scaling to the matrix $\mathbf{S}$, Bai and Yin (1988) derived the weak limit of the ESD of the transformed matrix when $\frac{p}{n} \to 0$.

Theorem 3 (Bai and Yin, 1988)**.** *Suppose that $\mathbf{X}$ is a $p \times n$ matrix with i.i.d. real-valued*

entries with mean 0 and variance 1 with finite fourth moment. Suppose $\lim_{n\to\infty} \frac{p}{n} = 0$ *. Then, as* $p \to \infty$*, the empirical spectral distribution (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 a nonrandom distribution, known as semi-circular distribution having the pdf*

$$f(x) = \frac{1}{2\pi}\sqrt{4 - x^2}, \quad -2 \leq x \leq 2 \tag{3.2}$$

Like the Marcenko-Pastur law, the above theorem is also distribution-free and provides the rate at which the eigenvalues of $\mathbf{S} = \frac{1}{n}\mathbf{X}\mathbf{X}^T$ goes to 1 when $\frac{p}{n} \to 0$. However, one potential disadvantage of the above two theorems is the requirement of the i.i.d data. In both of the theorem, in the data matrix $\mathbf{X} = [X_1, \cdots, X_n]$, where each of the columns (X_i) represents a data point of dimension p, though each of the columns can be assumed to be independent, in practice it is very unlikely that the entries of a single data point in $\mathbb{R}^p$ will be mutually independent as well. Henceforth substantial progress has been made to generalize these results by relaxing the conditions of independence within the columns. If $Y_1, \cdots, Y_n$ are i.i.d. $p-$dimensional data points with covariance matrix $\mathbf{\Sigma}$, then $\mathbf{Y} = [Y_1, \cdots, Y_n] = \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{X}$, where $\mathbf{X}$ satisfies the conditions of Theorem 2 and 3. In this context, Silverstein (1995) develops the Marcenko-Pastur law for the ESD of $\frac{1}{n}\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{X}\mathbf{X}^T\mathbf{\Sigma}^{\frac{1}{2}}$ when $\frac{p}{n} \to \gamma \in (0, \infty)$ under the same conditions as of Theorem 2. For $\frac{p}{n} \to 0$, under the conditions of Theorem 3, Bao (2012) showed the ESD of $\sqrt{\frac{n}{p}}\left(\frac{1}{n}\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{X}\mathbf{X}^T\mathbf{\Sigma}^{\frac{1}{2}} - \mathbf{\Sigma}\right) = \sqrt{\frac{n}{p}}\mathbf{\Sigma}^{\frac{1}{2}}\left(\frac{1}{n}\mathbf{X}\mathbf{X}^T - \mathbf{I}\right)\mathbf{\Sigma}^{\frac{1}{2}}$ converges almost surely in distribution to a nonrandom distribution.

Further research has been done to develop similar results under different forms of dependence. For instance, Yin and Krishnaiah (1987) derived similar results when $X_1, \cdots, X_n$ are i.i.d from a spherically symmetric distribution. Hui and Pan (2010), Wei et al. (2016) considered the case when the data points come from a $m-$dependent process. Hofmann-Credner and Stolz (2008) and Friesen et al. (2013) assumed that the entries of the data matrix $\mathbf{X} = [X_1, \cdots, X_n]$ can be partitioned into independent subsets while allowing the

entries from the same subset to be dependent. Gotze and Tikhomirov (2006) replaced the independent assumption by a technical martingale-type condition. Yao (2012) develops a version of the Marcenko-Pastur law when $X_1, \cdots, X_n$ are independent and entries of each X_i comes from a linear time series process.

3.1.2 Properties of extreme eigenvalues

In the previous section, we have a detailed description of the limit of the ESD of random matrices. However, in many situations, it is important to know whether the sample eigenvalues of $\mathbf{S}$ (as in Theorem 2) lie inside the support of F_γ as well. For instance, in signal processing, pattern recognition, edge detection, and many other areas, the support of the LSD of the population covariance matrices consists of several disjoint pieces. So it is essential to know whether or not the LSD of the sample covariance matrices is also separated into the same number of disjoint pieces, and under what conditions this is true. Also, many statistics can be written as a function of the integrals of the ESD of the random matrix. For example, the determinant of the sample covariance matrix is very useful in wireless communication and signal processing (Paul and Aue (2014)) which can be written as

$$\det(\mathbf{A}) = \prod_{j=1}^{n} \lambda_j = \exp\left(n \int_0^\infty \log x \, F^{\mathbf{A}}(dx) \right) \tag{3.3}$$

So under the knowledge of the asymptotic distribution of the ESD, usually the Helly-Bray theorem (Billingsley (2013)) is used to obtain an approximation of the statistic. But often such functions are not bounded (e.g. the function in 3.3 is $logx$ which is unbounded. As a result, the LSD and Helly-Bray theorem cannot be used to approximate the statistics. This limitation reduces the usefulness of the LSD. However, in many cases, the supports of the LSDs are compact intervals. Still, this alone does not guarantee that the Helly-Bray theorem can be applied unless one also proves in addition that the extreme eigenvalues of the random matrix stay within certain bounded intervals. These examples demonstrate that knowledge

about the weak limit of the ESD is not sufficient. Furthermore, extreme eigenvalues of random matrices themselves occur naturally in many problems such as principal component analysis. Henceforth studies regarding the asymptotic properties of the extreme eigenvalues of random matrices are extremely important. Under the assumption of the finiteness of the fourth moment of the i.i.d. entries, Yin et al. (1984) proved that the maximum eigenvalue of $\mathbf{S}$ (as in Theorem 2) converges almost surely.

Theorem 4 (Yin et al. (1984))**.** *Suppose that* $\mathbf{X}$ *is a* $p \times n$ *matrix with i.i.d. real-valued entries with mean 0 and variance* σ^2 *and finite fourth moment. Suppose* $\lim_{n\to\infty} \frac{p}{n} = \gamma \in (0, \infty)$ *. Suppose* $\lambda_{max}(n)$ *is the maximum eigenvalue of the* $p \times p$ *random matrix* $\mathbf{S} = \frac{1}{n}\mathbf{X}\mathbf{X}^T$*. Then*

$$\lim_{n\to\infty} \lambda_{max}(n) = (1 + \sqrt{\gamma})^2\sigma^2 \quad a.s. \tag{3.4}$$

A similar result was developed in Bai and Yin (2008) for the smallest eigenvalue of $\mathbf{S}$ as well under the same set of assumptions as of Theorem 5 as well when $p < n$.

Theorem 5 (Bai and Yin (2008))**.** *Suppose that* $\mathbf{X}$ *is a* $p \times n$ *matrix with i.i.d. real-valued entries with mean 0 and variance* σ^2 *and finite fourth moment. Suppose* $\lim_{n\to\infty} \frac{p}{n} = \gamma \in (0, 1)$ *. Suppose* $\lambda_{min}(n)$ *is the smallest eigenvalue of the* $p \times p$ *random matrix* $\mathbf{S} = \frac{1}{n}\mathbf{X}\mathbf{X}^T$*. Then*

$$\lim_{n\to\infty} \lambda_{min}(n) = (1 - \sqrt{\gamma})^2\sigma^2 \quad a.s. \tag{3.5}$$

These results give an accurate idea of the asymptotic range of the eigenvalues of sample covariance matrices under very mild assumptions. However many classical tests in multivariate analysis consist of the largest eigenvalues of sample covariance matrices (eg. Roy's largest root test) which makes the asymptotic distributions of the maximum eigenvalue of special interest. In the celebrated paper Johnstone (2001), the limiting distribution of the largest eigenvalue of the sample covariance matrix was derived when the entries of the data matrix are i.i.d from the standard normal distribution. Suppose that $\mathbf{X} = ((X_{ij}))$ is an $p \times n$

matrix with entries are i.i.d. from standard normal distribution,

$$X_{ij} \sim N(0,1).$$

Let l_1 be the largest sample eigenvalue of the Wishart matrix $\mathbf{X}\mathbf{X}^T$. Define the centering and scaling constants as follows:

$$\mu_{n,p} = \left(\sqrt{n-1} + \sqrt{p}\right)^2 \tag{3.6}$$

$$\sigma_{n,p} = \left(\sqrt{n-1} + \sqrt{p}\right)\left(\frac{1}{\sqrt{n-1}} + \frac{1}{\sqrt{p}}\right)^{1/3} \tag{3.7}$$

The Tracy-Widom law of order 1 has the distribution function defined by:

$$F_1(s) = \exp\left(-\frac{1}{2}\int_s^\infty \left[q(x) + (x-s)q^2(x)\right] dx\right), \quad s \in \mathbb{R} \tag{3.8}$$

where $q(x)$ solves the nonlinear Painlevé II differential equation:

$$q''(x) = xq(x) + 2q^3(x) \tag{3.9}$$

with the asymptotic condition:

$$q(x) \sim \mathrm{Ai}(x) \quad \text{as} \quad x \to +\infty, \tag{3.10}$$

where $\mathrm{Ai}(x)$ denotes the Airy function. This distribution was found by Tracy and Widom (1996) as the limiting law of the largest eigenvalue of an $n \times n$ Gaussian symmetric matrix. In terms of these distributions, the asymptotic distribution of l_1 can be stated as follows,

Theorem 6 (Johnstone (2001)). *Suppose that $\mathbf{X} = ((X_{ij}))$ is an $p \times n$ matrix whose entries are i.i.d. from standard normal distribution i.e. $X_{ij} \overset{i.i.d.}{\sim} N(0,1)$. If $\frac{p}{n} \to \gamma \in (0,\infty)$, and l_1 denotes the highest eigenvalue of $\mathbf{X}\mathbf{X}^T$ then,*

$$\frac{l_1 - \mu_{n,p}}{\sigma_{n,p}} \xrightarrow{\mathcal{L}} W_1 \sim F_1 \tag{3.11}$$

where $\mu_{n,p}, \sigma_{n,p}$ are as in 3.6 and 3.7 respectively and W_1 is a random variable following Tracy-Widom distribution defined in 3.8.

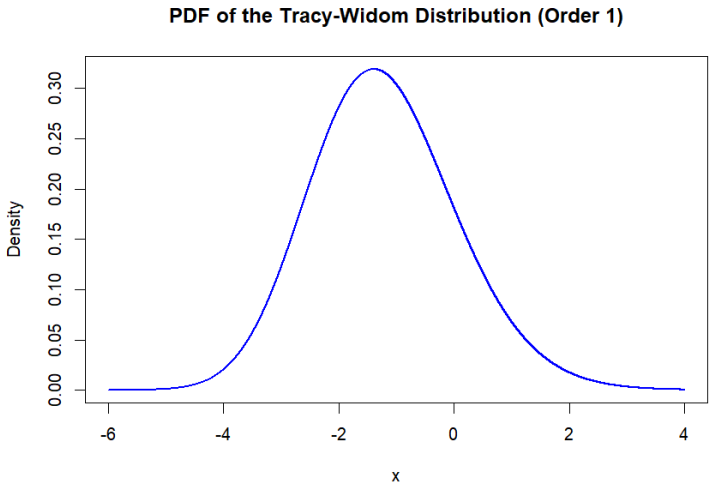


Figure 2: Density function for the Tracy-Widom Distribution

In Karoui (2003), the same result was extended for the cases $\gamma = 0, \infty$ as well. These results are very useful in single Wishart (e.g. principal component analysis (PCA), factor analysis, and tests for population covariance matrices in one-sample problems) and double Wishart problems (e.g. multivariate analysis of variance (MANOVA), canonical correlation analysis (CCA), tests for equality of covariance matrices and tests for linear hypotheses in multivariate linear regression problems). Many asymptotic generalizations of classical tests (e.g. Roy's largest root test) have been obtained using the above results which have a wide range of applications in signal processing, wireless communication (Paul and Aue (2014)), and machine learning (Couillet and Liao (2022)). Section 4 has a detailed discussion on these applications.

To check the practical applicability of Theorem 6, for the purpose of approximation, a simulation study was done (Johnstone (2001)). First, for square cases $n = p = 5, 10$ and 100, using $R = 10,000$ replications, results are shown in Table 1. Even for 5×5 and 10×10, the approximation seems to be quite good in the right-hand tail at conventional significance levels of 10%, 5%, and 1%. At 100×100, the approximation seems reasonable throughout the

range. The same general picture holds for n/p in the ratio 4:1. Even for 5×20 matrices, the approximation is reasonable, if not excellent, at the conventional upper significance levels.

A further summary message from these computations is that in the null Wishart case, about 80% of the distribution lies below μ_{np} and 95% below $\mu_{np} + \sigma_{np}$. Ma (2012) has a detailed discussion on the accuracy of the Tracy-Widom laws.

Table 1: (Johnstone (2001)) Simulations for finite $n \times p$ versus Tracy–Widom Limit. The first column shows the probabilities of the F_1 limit distribution corresponding to fractions in the second column. The next three columns show estimated cumulative probabilities for l_1, centered and scaled as in Equation (3.6) and 3.7, in $R = 10,000$ repeated draws from $W_p(n, I)$ with $n = p = 5, 10, 100$. The following three cases have $n : p$ in the ratio 4:1. The final column gives approximate standard errors based on binomial sampling. The bold font highlights some conventional significance levels. The Tracy–Widom distribution F_1 was evaluated on a grid of 121 points $-6(0.1)6$ using the Mathematica package `p2Num` written by Craig Tracy. The remaining computations were done in MATLAB, with percentiles obtained by inverse interpolation and using `randn()` for normal variates and `norm()` to evaluate the largest singular values.

Percentile	TW	5×5	10×10	100×100	5×20	10×40	100×400	$2 \times$ SE
-3.90	0.01	0.000	0.001	0.007	0.002	0.003	0.010	(0.002)
-3.18	0.05	0.003	0.015	0.042	0.029	0.039	0.049	(0.004)
-2.78	0.10	0.019	0.049	0.089	0.075	0.089	0.102	(0.006)
-1.91	0.30	0.211	0.251	0.299	0.304	0.307	0.303	(0.009)
-1.27	0.50	0.458	0.480	0.500	0.539	0.524	0.508	(0.010)
-0.59	0.70	0.697	0.707	0.703	0.739	0.733	0.714	(0.009)
0.45	0.90	0.901	0.907	0.903	0.919	0.918	0.908	(0.006)
0.98	0.95	0.948	0.954	0.950	0.960	0.961	0.957	(0.004)
2.02	0.99	0.988	0.991	0.991	0.992	0.993	0.992	(0.002)

3.2 Asymptotic Properties of $F-$type matrices

In this section, we discuss the asymptotic properties of a multivariate $F-$ matrix. Multivariate F-distribution plays a crucial role in several areas of multivariate data analysis, especially when the relationships between multiple variables are tested simultaneously. It has primary application in two-sample tests on covariance matrices, MANOVA (multivariate analysis of

variance), multivariate linear regression, and in Canonical Correlation Analysis. Pioneering work by Wachter (1980) examined the limiting distribution of the solutions to the equation:

$$\det\left(\mathbf{X}_{1,n_1}\mathbf{X}_{1,n_1}^T - \lambda\mathbf{X}_{2,n_2}\mathbf{X}_{2,n_2}^T\right) = 0,$$

where $\mathbf{X}_{j,n_j}$ is a $p \times n_j$ matrix with i.i.d. entries from $N(0,1)$, and $\mathbf{X}_{1,n_1}$ is independent of $\mathbf{X}_{2,n_2}$. When $\mathbf{X}_{2,n_2}\mathbf{X}_{2,n_2}^T$ is of full rank, the solutions to this equation are $\frac{n_2}{n_1}$ times the eigenvalues of the multivariate F-matrix:

$$\left(\frac{1}{n_1}\mathbf{X}_{1,n_1}\mathbf{X}_{1,n_1}^T\right)\left(\frac{1}{n_2}\mathbf{X}_{2,n_2}\mathbf{X}_{2,n_2}^T\right)^{-1}.$$

Yin and Krishnaiah (1983) proved the existence of the limiting spectral distribution (LSD) of the matrix sequence $\{\mathbf{S}_n\mathbf{T}_n\}$, where $\mathbf{S}_n$ is a standard Wishart matrix of dimension p with n degrees of freedom, and $\frac{p}{n} \to \gamma \in (0,\infty)$, $\mathbf{T}_n$ is a positive definite matrix with $\beta_k(\mathbf{T}_n) \to \mathbf{H}_k$, and the sequence $\mathbf{H}_k$ satisfies the Carleman condition. In Yin (1986), this result was extended to the case where the sample covariance matrix is formed from i.i.d. real random variables with mean zero and variance one. Building on the work of Yin and Krishnaiah (1983), later Yin et al. (1983) demonstrated the existence of the LSD of the multivariate F-matrix. The explicit form of the LSD for multivariate F-matrices was derived by Bai et al. (1988) and Silverstein (1995) and is given by the following theorem.

Theorem 7 (Bai et al. (1988))**.** *Let* $\mathbf{F} = \mathbf{S}_{n_1}\mathbf{S}_{n_2}^{-1}$*, where* $\mathbf{S}_{n_i}$ *(for* $i = 1,2$*) is a sample covariance matrix with dimension* p *and sample size* n_i*, and the underlying distribution has mean 0 and variance 1. If* $\mathbf{S}_{n_1}$ *and* $\mathbf{S}_{n_2}$ *are independent,* $\frac{p}{n_1} \to \gamma \in (0,\infty)$*, and* $\frac{p}{n_2} \to \gamma' \in (0,1)$*, then the limiting spectral distribution (LSD)* $F_{\gamma,\gamma'}$ *of* $\mathbf{F}$ *exists and has a density function given by:*

$$f_{\gamma,\gamma'}(x) = \begin{cases} \frac{(1-\gamma')\sqrt{(b-x)(x-a)}}{2\pi x(\gamma+x\gamma')}, & \text{if } a < x < b, \\ 0, & \text{otherwise,} \end{cases}$$

where

$$a = \left(\frac{1-\sqrt{\gamma+\gamma'-\gamma\gamma'}}{1-\gamma'}\right)^2, b = \left(\frac{1+\sqrt{\gamma+\gamma'-\gamma\gamma'}}{1-\gamma'}\right)^2.$$

Further, if $\gamma > 1$*, then* $F_{\gamma,\gamma'}$ *has a point mass* $1 - \frac{1}{\gamma}$ *at the origin.*

Besides the entire ESD, the extreme eigenvalues of multivariate $F-$matrices are also immensely important in many high-dimensional problems such as testing sphericality in covariance matrices, testing equality of multiple covariance matrices, correlated noise detection, etc (Han et al. (2016)). The following result from the phenomenal work of Han et al. (2016) obtains the limiting distribution of the generalized $F-$type matrices under mild assumptions. Before starting the actual theorem, we first state a condition the data matrix needs to satisfy.

Definition 1. *A real random matrix* $\mathbf{Z}$ *is said to satisfy* ***Condition 1****, if it consists of entries* $\{Z_{ij}\}$ *where* $\{Z_{ij}\}$ *are independent random variables with* $\mathbb{E}[Z_{ij}] = 0$ *and* $\mathbb{E}[|Z_{ij}|^2] = 1$ *and for all* $k \in \mathbb{N}$*, there exists a constant* C_k *such that* $\mathbb{E}[|Z_{ij}|^k] \leq C_k$*.*

In conjunction with the above definition, the following theorem presents the desired limiting distribution.

Theorem 8 (Han et al. (2016))**.** *Also assume that the real random matrices* $\mathbf{X} = (X_{ij})_{p\times n}$ *and* $\mathbf{Y} = (Y_{ij})_{p\times m}$ *are independent and satisfies* ***Condition 1****. Set* $m = m(p)$ *and* $n = n(p)$*. Suppose that*

$$\lim_{p\to\infty}\frac{p}{m} = d_1 > 0, \quad \lim_{p\to\infty}\frac{p}{n} = d_2 > 0, \quad 0 < \lim_{p\to\infty}\frac{p}{m+n} < 1.$$

satisfies $0 < d_1 < 1$ *and,* $0 < d_2 < \infty$*. Let,*

$$\breve{m} = \max\{m, p\}, \quad \breve{n} = \min\{n, m+n-p\}, \quad \breve{p} = \min\{m, p\}.$$

Moreover, let

$$\sin^2\left(\frac{\gamma}{2}\right) = \frac{\min\{\breve{p},\breve{n}\} - \frac{1}{2}}{\breve{m}+\breve{n}-1}, \quad \sin^2\left(\frac{\psi}{2}\right) = \frac{\max\{\breve{p},\breve{n}\} - \frac{1}{2}}{\breve{m}+\breve{n}-1},$$

$$\mu_{J,p} = \tan^2\left(\frac{\gamma+\psi}{2}\right), \tag{3.12}$$

$$\sigma^3_{J,p} = \frac{16\mu^3_{J,p}}{(\breve{m}+\breve{n}-1)^2} \cdot \frac{1}{\sin(\gamma)\sin(\psi)\sin^2(\gamma+\psi)}. \tag{3.13}$$

Set

$$\mathbf{B}_p = \frac{\mathbf{X}\mathbf{X}^T}{\breve{n}} \quad \textit{and} \quad \mathbf{A}_p = \frac{\mathbf{Y}\mathbf{Y}^T}{\breve{m}}.$$

Denote the largest root of

$$\det(\lambda\mathbf{A}_p - \mathbf{B}_p) = 0$$

by λ_1. *Then*

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

where $F_1(s)$ *is the cumulative distribution function of the Tracy-Widom distribution defined in Equation* (3.8).

The above theorem has a couple of interesting remarks. For instance, this immediately implies the distribution of the largest root of $\det(\lambda(\mathbf{B}_p + \mathbf{A}_p) - \mathbf{B}_p) = 0$. In fact, the largest root of $\det(\lambda(\mathbf{B}_p + \mathbf{A}_p) - \mathbf{B}_p) = 0$ is $\frac{\lambda_1}{1+\lambda_1}$ if λ_1 is the largest root of the F matrices $\mathbf{B}_p\mathbf{A}_p^{-1}$ in Theorem 8 when $0 < d_1 < 1$.

When $d_1 > 1$, the largest root of $\det(\lambda(\mathbf{B}_p+\mathbf{A}_p)-\mathbf{B}_p) = 0$ is one with multiplicity $(p-m)$. In that case, instead one considers the $(p-m+1)$th largest root of $\det(\lambda(\mathbf{B}_p+\mathbf{A}_p)-\mathbf{B}_p) = 0$. It turns out that the $(p-m+1)$th largest root of $\det(\lambda(\mathbf{B}_p+\mathbf{A}_p)-\mathbf{B}_p) = 0$ is $\frac{\lambda_1}{1+\lambda_1}$ if λ_1 is the largest root of $\det(\lambda\mathbf{A}_p - \mathbf{B}_p) = 0$. The exact order of the centering and scaling parameters $\mu_{J,p}$ and $\sigma_{J,p}$ can also be obtained along the lines of this result in terms of $\breve{m}, \breve{n}$ and p. Section 4 has a detailed discussion on the applications of these results on high-dimensional inference.

4 Applications in Statistics

In this section, we discuss various applications of random matrix theory in statistics and related fields. So far there are a lot of ground-breaking applications of RMT that helped to develop robust, efficient high dimensional data handling methods, enriched complex machine learning algorithms, optimized signal processing techniques and motivated a lot of crucial discoveries in genomics, finance, climate science, and social network analysis. Henceforth, in this section, the key focus is on these applications to real-world problems in conjunction with the theoretical discussion above.

4.1 Inference on Covariance Matrices

One of the primary applications of the theory of large random matrices in high-dimensional statistics is inference on covariance matrices. Since the results provide asymptotic properties of the spectra of large random matrices, using those results one can check whether one estimate is consistent under certain conditions, and also for hypothesis testing, one can approximate the power functions if the test statistic is a function of the spectrum. In this regard, One of the earliest uses of the distribution of the largest eigenvalue of the sample covariance matrix is in testing the hypothesis $H_0 : \Sigma = \mathbf{I}_p$ when i.i.d. samples are drawn from a $N(\mu, \Sigma)$ distribution. The Tracy–Widom law for the largest sample eigenvalue under the null Wishart case, i.e., when the population covariance matrix $\Sigma = \mathbf{I}_p$, allows a precise determination of the cut-off value for this test, which, with a careful calibration of the centering and normalizing sequences, is very accurate even for relatively small p and n (Johnstone (2001),Johnstone and Lu (2009)).

The behavior of the power of the test requires formulating suitable alternative models. For instance, for data matrix $\mathbf{X} = [X_1, \cdots, X_p] \in \mathbb{R}^{p\times n}$ with iid columns x_i, consider the

testing problem,

$$H_0 : \mathbf{X} = \sigma \mathbf{Z}$$
$$H_1 : \mathbf{X} = a\mathbf{s}^T + \sigma \mathbf{Z} \tag{4.1}$$

where $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n] \in \mathbb{R}^{p \times n}$ with $\mathbf{z}_i \sim N(\mathbf{0}, \mathbf{I}_p)$, $\mathbf{a} \in \mathbb{R}^p$ deterministic with unit norm $\|\mathbf{a}\| = 1$, $\mathbf{s} = [s_1, \ldots, s_n]^\top \in \mathbb{R}^n$ with s_i i.i.d. random scalars, and $\sigma > 0$. We also denote $c = p/n$ (and demand as usual that $0 < \liminf c \leq \limsup c < \infty$).

This model describes the observation of either pure Gaussian noise data $\sigma \mathbf{z}_i$ with zero mean and covariance $\sigma^2 \mathbf{I}_p$, or of deterministic information $\mathbf{a}$ possibly modulated by a scalar (random) signal s_i (which could simply be ± 1) added to the noise. If the parameters $\mathbf{a}$, σ as well as the statistics of s_i are known, a mere Neyman-Pearson test allows one to discriminate between H_0 and H_1 with optimal detection probability, for all finite n, p; precisely, one will decide on the genuine hypothesis according to the ratio of posterior probabilities

$$\frac{\mathbb{P}(\mathbf{X} \mid H_1)}{\mathbb{P}(\mathbf{X} \mid H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \alpha \tag{4.2}$$

for some $\alpha > 0$ controlling the desired Type I and Type II error rates (that is, the probability of false positives and of false negatives).

However, in practice, unless the existence of a set of previous pure-noise acquisitions is assumed, it is quite unlikely that σ be assumed known or consistently estimated. Similarly, if the ultimate objective (post-decision) is to estimate the data structure $\mathbf{a}$ under H_1, $\mathbf{a}$ is naturally assumed partially or completely unknown (it may be known to belong to a subset of $\mathbb{R}^p$ in which case more elaborate procedures than proposed here can be carried on). In the most generic scenario where $\mathbf{a}$ is fully unknown, assuming additionally the data of zero mean, we may thus impose without generality the restriction that Under this (very restricted) prior knowledge, instead of the maximum likelihood test in (4.2), one may resort to a *generalized likelihood ratio test (GLRT)* defined as

$$\frac{\sup_{\sigma, \mathbf{a}} \mathbb{P}(\mathbf{X} \mid \sigma, \mathbf{a}, \mathcal{H}_1)}{\sup_{\sigma, \mathbf{a}} \mathbb{P}(\mathbf{X} \mid \sigma, \mathbf{a}, \mathcal{H}_0)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \alpha.$$

Under both Gaussian noise and signal s_i assumption, the GLRT has an explicit expression that appears to be a monotonously increasing function of $\|\mathbf{X}\mathbf{X}^\top\|/\mathrm{tr}(\mathbf{X}\mathbf{X}^\top)$. That is, the test is equivalent to

$$T_p \equiv \frac{\|\frac{1}{n}\mathbf{X}\mathbf{X}^\top\|}{\frac{1}{p}\mathrm{tr}\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\top\right)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} f(\alpha),$$

(Wax and Kailath (1985) and Anderson (1963) has a detailed discussion on this idea) for some known monotonously increasing function f. Here we introduced the normalizations $1/p$ and $1/n$ so that both the numerator and denominator are of order $O(1)$ as $n, p \to \infty$.

Since the ratio T_p has limit $(1+\sqrt{c})^2$ under the H_0 asymptotics, $f(\alpha)$ must be of the form $f(\alpha) = (1+\sqrt{c})^2 + g(\alpha)$ for some $g(\alpha) > 0$. Also, as we know that $\frac{1}{p}\mathrm{tr}\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\top\right)$ fluctuates at the speed $O(n^{-1})$, while $\|\frac{1}{n}\mathbf{X}\mathbf{X}^\top\|$ fluctuates at the slower speed $O(n^{-2/3})$ (as per Theorem 6), the global fluctuation is dominated by the numerator at a rate of order $O(n^{-2/3})$, i.e., we have under H_0,

$$T_p \overset{H_0}{=} (1+\sqrt{c})^2 + O(n^{-2/3}).$$

Since the denominator essentially converges (at an $O(n^{-1})$ rate) while the numerator still fluctuates (at an $O(n^{-2/3})$ rate), despite the dependence between both, only the fluctuations of the numerator $\frac{1}{n}\mathbf{X}\mathbf{X}^\top$ influence the behavior of the ratio T_p, and thus

$$T_p \overset{H_0}{\sim} (1+\sqrt{c})^2 + (1+\sqrt{c})\frac{4}{3}c^{-\frac{1}{6}}n^{-\frac{2}{3}}\mathrm{TW} + o(n^{-2/3}),$$

where TW denotes the Tracy-Widom Distribution. As a consequence, in order to set a maximum false alarm rate (or false positive, or Type I error) of $r > 0$ in the limit of large n, p, one must choose a threshold $f(\alpha)$ for T_p such that

$$\mathbb{P}(T_p \geq f(\alpha)) = r,$$

that is, such that

$$\mu_{\mathrm{TW}}([A_p, +\infty)) = r, \quad A_p = (f(\alpha) - (1+\sqrt{c})^2)(1+\sqrt{c})^{-\frac{4}{3}}c^{\frac{1}{6}}n^{\frac{2}{3}} \tag{3.2}$$

with μ_{TW}, the Tracy-Widom measure.

For testing problems on covariance matrices, with a both-sided alternative, based on a normal random sample, instead of Tracy-Widom law, one can also use Theorem 1. For $X_1, \cdots, X_n \overset{iid}{\sim} N_p(\mu, \mathbf{\Sigma})$ where μ, Σ are unknown and p, n are both large and, $p \sim n^{\frac{1}{2}-\epsilon}$, $\epsilon > 0$. Consider the testing problem with a two-sided alternative,

$$\begin{aligned} H_0 &: \mathbf{\Sigma} = \mathbf{\Sigma}_0 \\ H_1 &: \mathbf{\Sigma} \neq \mathbf{\Sigma}_0 \end{aligned} \tag{4.3}$$

From Theorem 1, it turns out

$$\phi(\mathbf{X}) = \mathbf{1}\left(\sqrt{\frac{n-1}{2p}}\left|\sum_{i=1}^{p}\log\left(\frac{\hat{\lambda}_i}{\lambda_i}\right) - \sum_{i=1}^{p}\log(n-p+i)\right| > z_{\alpha/2}\right) \tag{4.4}$$

is an asymptotically size α test, where $\hat{\lambda}_1, \cdots, \hat{\lambda}_p$ are the eigenvalues of the sample covariance matrix $S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)(X_i - \overline{X}_n)^T$, $\lambda_1, \cdots, \lambda_p$ are the eigenvalues of $\mathbf{\Sigma}_0$, $\alpha \in (0,1)$ and z_α is the $(1-\alpha)$th quantile of N(0,1).

The same idea can be generalized for testing problems of the covariance matrices in high-dimensional Linear Regression when the number of response variables and number of data points are both large. Consider the multivariate Linear Regression model,

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{4.5}$$

where $\mathbf{Y} = [Y_1 : \cdots : Y_m] \in \mathbb{R}^{n\times m}$ 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. For inference purposes, it is further assumed that $\mathbf{E} \sim NDM(0, \mathbf{\Sigma})$ i.e. rows of $\mathbf{E}$ are iid from $N(0, \mathbf{\Sigma})$. Consider the testing problem with a two-sided alternative, as in (4.3) i.e.

$$\begin{aligned} H_0 &: \mathbf{\Sigma} = \mathbf{\Sigma}_0 \\ H_1 &: \mathbf{\Sigma} \neq \mathbf{\Sigma}_0 \end{aligned}$$

Under H_0, the sum of squares of error (SSE) defined as $\mathbf{Y}^T(\mathbf{I}-\mathbf{P_X})\mathbf{Y}$, where $\mathbf{P_X}$ is the orthogonal projection matrix of $\mathcal{C}(\mathbf{X})$ follows $W_m(\mathbf{\Sigma}_0, n-r)$ with r to be the rank of $\mathbf{X}$. Therefore an asymptotically level α test to test (4.3) is given by

$$\phi_{\mathcal{R}} := \mathbf{1}\left(\sqrt{\frac{n-r}{2m}}\left|\sum_{i=1}^{m}\log\left(\frac{\hat{\lambda}_i}{\lambda_i}\right) - \sum_{i=1}^{m}\log(n-r-k+i)\right| > z_{\alpha/2}\right) \tag{4.6}$$

where $\hat{\lambda}_1, \cdots, \hat{\lambda}_p$ are the eigenvalues of SSE $= \mathbf{Y}^T(\mathbf{I}-\mathbf{P_X})\mathbf{Y}$, $\lambda_1, \cdots, \lambda_p$ are the eigenvalues of $\mathbf{\Sigma}_0$, $\alpha \in (0,1)$ and z_α is the $(1-\alpha)$th quantile of N(0,1).

The same idea can be generalized for the two sample tests of equality for covariance matrices as well. Suppose we have data points $X_1, \cdots, X_m \overset{iid}{\sim} N_p(\mu_1, \mathbf{\Sigma}_1)$ and $Y_1, \cdots, Y_n \overset{iid}{\sim} N_p(\mu_2, \mathbf{\Sigma}_2)$ where $\mu_1, \mu_2, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2$ are unknown. $n \equiv n(m)$ satisfies $\lim_{m\to\infty}\frac{n}{m} = c \in (0,\infty)$. Consider the testing problem,

$$\begin{aligned} H_0 &: \mathbf{\Sigma_1} = \mathbf{\Sigma_2} \\ H_1 &: \mathbf{\Sigma_1} \neq \mathbf{\Sigma_2} \end{aligned} \tag{4.7}$$

Then by Theorem 1 it turns out that

$$\phi(\mathbf{X},\mathbf{Y}) := \mathbf{1}\left(\sqrt{\frac{m}{2p\left(1+\frac{1}{c}\right)}}\left|\sum_{i=1}^{p}\log\left(\frac{\hat{\lambda}_i}{\hat{\lambda}_i^*}\right) - \sum_{i=1}^{p}\log\left(\frac{n-p+i}{m-p+i}\right)\right| > z_{\alpha/2}\right) \tag{4.8}$$

is an asymptotically level α test where $\hat{\lambda}_1, \cdots, \hat{\lambda}_p$ are the eigenvalues of the sample covariance matrix $\mathbf{S_x} = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \overline{X}_m)(X_i-\overline{X}_m)^T$, $\hat{\lambda}_1^*, \cdots, \hat{\lambda}_p^*$ are the eigenvalues of $\mathbf{S_Y} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\overline{Y}_n)(Y_i-\overline{Y}_n)^T$, $\alpha \in (0,1)$ and z_α is the $(1-\alpha)$th quantile of N(0,1). The power function for this test is given by, $\beta(\mathbf{\Sigma_1}, \mathbf{\Sigma_2}) := 1 - \Phi\left(z_{\alpha/2} - \sqrt{\frac{m}{2p\left(1+\frac{1}{c}\right)}}\sum_{i=1}^{p}\log\left(\frac{\lambda_i}{\lambda_i^*}\right)\right) + \Phi\left(-z_{\alpha/2} - \sqrt{\frac{m}{2p\left(1+\frac{1}{c}\right)}}\sum_{i=1}^{p}\log\left(\frac{\lambda_i}{\lambda_i^*}\right)\right) + f(n,m)$ where $f(n,m) = O\left(p\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right)\right)$.

The same test can be done using the asymptotic theory of the largest root of $F-$type matrices as well. Since $\mathbf{S_Y}$ is almost surely invertible, we can define the test,

$$\phi_{\mathcal{F}} := \mathbf{1}\left(\frac{\frac{\breve{n}}{\breve{m}}\lambda_1 - \mu_{J,p}}{\sigma_{J,p}} > F_1^{-1}(1-\alpha)\right) \tag{4.9}$$

where $\breve{n}, \breve{m}, \mu_{J,p}, \sigma_{J,p}, F_1(\cdot)$ are as in Theorem 8 and λ_1 is the largest root of

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

By Theorem 8, $\phi_{\mathcal{F}}$ turns out to be an asymptotic size α test.

For the high-dimensional linear regression model defined in (4.5), one can also obtain a high-dimensional generalization of Wald's test using the asymptotic theories of the $F-$type matrices. Consider the wald's testing problem

$$\begin{aligned} H_0 &: \mathbf{L^T B} = \mathbf{B_0} \\ H_1 &: \mathbf{L^T B} \neq \mathbf{B_0} \end{aligned} \tag{4.10}$$

where $\mathbf{L} \in \mathbb{R}^{p\times k}$ and rank of $\mathbf{L}$ is k. If the design matrix $\mathbf{X}$ has full column rank, i.e. $\rho(\mathbf{X}) = p$, the Ordinary Least Squares (OLS) estimate for $\mathbf{B}$ is given by $\hat{\mathbf{B}} = (\mathbf{X^T X})^{-1}\mathbf{X^T Y}$. Then under H_0,

$$\mathbf{A_p} := (\mathbf{L^T}\hat{\mathbf{B}} - \mathbf{B_0})^{\mathbf{T}}(\mathbf{L^T}(\mathbf{X^T X})^{-1}\mathbf{L})^{-1}(\mathbf{L^T}\hat{\mathbf{B}} - \mathbf{B_0}) \overset{H_o}{\sim} W_m(\mathbf{\Sigma}, k) \tag{4.11}$$

and

$$\mathbf{B_p} = \mathbf{Y^T}(\mathbf{I} - \mathbf{P_X})\mathbf{Y} \sim W_m(\mathbf{\Sigma}, n-p) \tag{4.12}$$

and $\mathbf{A_p}, \mathbf{B_p}$ are independent. Let λ_1 be the largest root of $det(\lambda A_p - B_p)$. Then from Theorem 8, it turns out that a normalized version of λ_1 asymptotically follows Tracy-Widom Distribution under H_0. Therefore, in view of Theorem 8, one can construct an asymptotic size α test using λ_1.

4.2 Application in PCA

In Section 3.1 we discussed the LSD and asymptotic properties of the sample covariance matrix under the Gaussianity assumption when the eigenvalues of the population covariance matrix are either identical or are evenly spread out so that none of them "sticks out" from

the bulk. Soshnikov (2002) proved the distributional limits under weaker assumptions, in addition to deriving distributional limits of the $k-$th largest eigenvalue, for fixed but arbitrary k. Under the Gaussianity assumption of the data, the asymptotic distribution of the eigenvalues of the sample covariance matrix also turns out to be Gaussian if the eigenvalues of the population covariance matrix are distinct.

Theorem 9 (Mardia et al. (2024))**.** *Let* $X_1, \cdots, X_n \overset{iid}{\sim} N_p(0, \mathbf{\Sigma})$ *where* $\mathbf{\Sigma}$ *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* $\mathbf{\Lambda} = diag(\lambda_1, \cdots, \lambda_p)$*. Then*

$$\sqrt{n}(l_n - \lambda) \xrightarrow{\mathcal{L}} N_p(0, 2\mathbf{\Lambda}^2) \tag{4.13}$$

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

So the consistency of the sample eigenvalues of the sample covariance matrix holds when the population covariance matrix has either all eigenvalues identical or all distinct from each other. However, in recent years, researchers in various fields have been using different versions of covariance matrices of growing dimensions with special patterns. For instance, in speech recognition (Hastie et al. (1995)), wireless communication (Telatar (1999)), and statistical learning (Hoyle and Rattray (2003)) a few of the sample eigenvalues have limiting behavior that is different from the behavior when the covariance is the identity.

While high-dimensional data often exhibits complex patterns, it's frequently characterized by a simple underlying structure. This structure can be modeled as a low-dimensional "signal" obscured by high-dimensional "noise." Assuming an additive relationship between these components, we can represent the data using a factor model. Factor models are particularly useful for detecting and estimating low-dimensional signals within isotropic or nearly isotropic noise. Key statistical questions, such as those related to dimension reduction, can

be effectively addressed by analyzing the eigenvalues and eigenvectors of the sample covariance matrix. A particularly useful idealized model of this kind, named the spiked covariance model by Johnstone (2001) has been in use for quite some time in statistics. Under this model, the population covariance matrix $\mathbf{\Sigma}$ is expressed as

$$\mathbf{\Sigma} = \sum_{j=1}^{M} \lambda_j \theta_j \theta_j^* + \sigma^2 I_p, \tag{4.14}$$

where $\theta_1, ..., \theta_M$ are orthonormal; $\lambda_1 \geq ... \geq \lambda_M > 0$ and $\sigma^2 > 0$. This model implies that, except for M leading eigenvalues $l_j = \lambda_j + \sigma^2$ for $j = 1, ..., M$, the rest of the eigenvalues are all equal.

This model has been studied extensively in the context of high-dimensional PCA since it brings out several key issues associated with dimension reduction in the high-dimensional context. Johnstone and Lu (2009) first demonstrated that if $\frac{p}{n} \to \gamma \in (0, \infty)$ the sample principal components are inconsistent estimates of the population principal components under (4.14). This phase transition phenomenon is described in its simplest form in the following theorem, where, for convenience, we assume in (4.14) $\sigma^2 = 1$.

Theorem 10 (Baik and Silverstein (2006))**.** *Suppose that Σ is a $p \times p$ positive definite matrix with eigenvalues $\ell_1 \geq \cdots \geq \ell_M > 1 = \cdots = 1$, and let $\hat{\ell}_1 \geq \cdots \geq \hat{\ell}_p$ be the eigenvalues of the sample covariance matrix $S = n^{-1}\Sigma^{1/2} Z Z^* \Sigma^{1/2}$ where the $p \times n$ data matrix Z has i.i.d. real or complex entries with zero mean, unit variance and finite fourth moment. Suppose that $p, n \to \infty$ such that $p/n \to \gamma \in (0, \infty)$. Then, for each fixed $j = 1, 2, \cdots, M$*

$$\hat{\ell}_j \xrightarrow{a.s.} \begin{cases} (1 + \sqrt{\gamma})^2 & \text{if } \ell_j \leq 1 + \sqrt{\gamma}, \\ \ell_j \left(1 + \frac{\gamma}{\ell_j - 1}\right) & \text{if } \ell_j > 1 + \sqrt{\gamma}. \end{cases} \tag{4.15}$$

Therefore, when the population covariance matrix is of the spike form, it might not be such a good idea to use Principal Component Analysis (PCA) for dimension reduction in a high-dimensional setting, at least not in its standard form. In this regard, one natural

question is how one can test if the population covariance matrix $\mathbf{\Sigma}$ is in the form of (4.14) and how bad the inconsistency of the sample principal components if $\mathbf{\Sigma}$ is in spike form. The following theorem provides an answer to these questions.

Theorem 11 (Paul (2007)). *Suppose that* $X_1, \cdots, X_n \overset{iid}{\sim} N_p(0, \Sigma)$ *where* Σ *is a* $p \times p$ *positive definite matrix with eigenvalues* $\ell_1 \geq \cdots \geq \ell_M > 1 = \cdots = 1$, *and let* $\hat{\ell}_1 \geq \cdots \geq \hat{\ell}_p$ *be the eigenvalues of the sample covariance matrix* $S = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T$. *Suppose that* $p, n \to \infty$ *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*

$$\sqrt{n}\left(\hat{\ell}_j - \ell_j\left(1 + \frac{\gamma}{\ell_j - 1}\right)\right) \xrightarrow{\mathcal{L}} N(0, \sigma^2(\ell_j)) \tag{4.16}$$

as $n \to \infty$ *where* $\sigma^2(\ell) := 2\ell^2\left(1 - \frac{\gamma}{(\ell-1)^2}\right)$

Suppose that we test the hypothesis $H_0 : \Sigma = I$ versus the alternative that $H_1 : \Sigma = \text{diag}(\ell_1, \ldots, \ell_M, 1, \ldots, 1)$ with $\ell_1 \geq \cdots \geq \ell_M > 1$, based on i.i.d. observations from $N(0, \Sigma)$. If $\ell_1 > 1 + \sqrt{\gamma}$, it follows from Theorem 10 that the largest root test is asymptotically consistent. For the special case when ℓ_1 is of multiplicity one, Theorem 11 gives an expression for the asymptotic power function, assuming that p/n converges to γ fast enough, as $n \to \infty$. One has to view this in context since the result is derived under the assumption that $\ell_1, \ldots, \ell_M$ are all fixed, and we do not have a rate of convergence for the distribution of $\hat{\ell}_1$ toward normality. However, Theorem 11 can be used to find confidence intervals for the larger eigenvalues under the non-null model.

Under the same set of assumptions as of Theorem 11, Paul (2007) proved further that, if $\ell_j \leqslant 1 + \sqrt{\gamma}$ and ℓ_j is of arithmetic multiplicity one, then the angle between the $j-$th sample and population eigenvectors converges to $\frac{\pi}{2}$ almost surely which essentially shows in which extent the sample principal components can be inconsistent and provides a generalization of Johnstone and Lu (2009). Later on Bai and Zhou (2008) extends the results of Paul (2007) in the context of spiked covariance matrix by dropping the Gaussianity assumption.

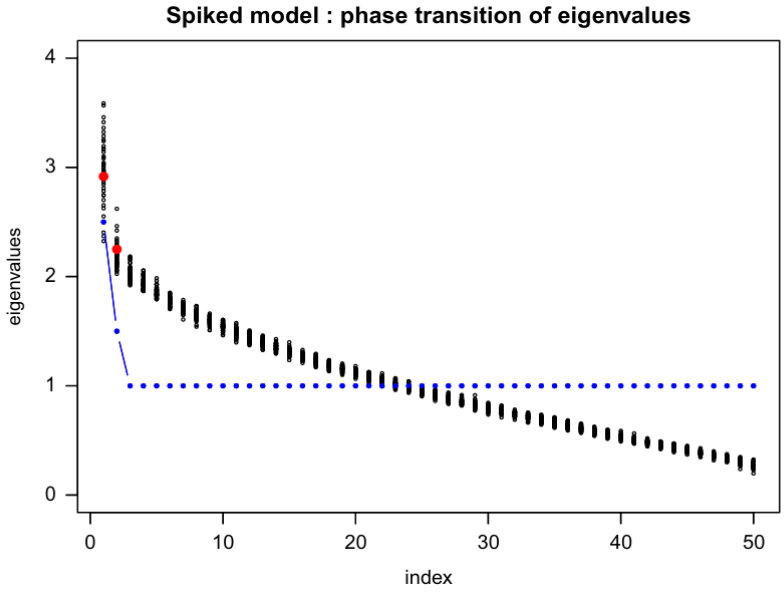


Figure 3: (Paul (2007)) An illustration of the phase transition of eigenvalues in a spiked covariance model: here, $p = 50$, $n = 200$ and eigenvalues of the covariance matrix are $\ell_1 = 2.5$, $\ell_2 = 1.5$, $\ell_j = 1$ for $j = 3, \ldots, p$. So, $\ell_1 > 1 + \sqrt{p/n}$ and $\ell_2 = 1 + \sqrt{p/n}$. 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 $\gamma = p/n = 0.25$.

4.3 On signal processing and wireless communications

Large random matrices come up often in signal processing, especially in wireless communication. Bai and Silverstein (2010), Couillet and Debbah (2011), and Tulino et al. (2004) highlight several such cases, including: (i) finding the channel capacity of MIMO (multiple-input-multiple-output) systems, which involves calculating the logarithm of the determinant of the matrix $\mathbf{I} + \mathbf{S}$, where $\mathbf{S}$ is a Wishart matrix that reflects the signal-to-noise ratio in transmission; (ii) finding the limiting SINR (signal-to-interference-noise ratio) in random channels and in linearly precoded systems, like CDMA (code-division-multiple-access) systems (Bai and Silverstein (2007)); (iii) analyzing the performance of receivers as the system size grows; and (iv) estimating energy from multiple sources (Couillet and Debbah (2011)). Besides, random matrices are useful in a variety of signal-processing problems, such as detect-

ing input signals (Nadakuditi and Silverstein (2010); Silverstein and Combettes (1992)) and estimating subspaces in sensor networks (Hachem et al. (2013)). Furthermore, the asymptotic distribution of the spectra of large random matrices and the idea behind Roy's largest root test (Roy (1953), Johnstone and Nadler (2017)) can be used to construct nonparametric tests to detect the number of signals embedded in noise (Kritchman and Nadler (2009)).

The standard setup for signals impinging on an array with sensors consists of n i.i.d $p-$dimensional observations $\{\mathbf{x}_i\}_{i=1}^n$ from the model,

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \sigma\mathbf{n}(t) \tag{4.17}$$

sampled at n distinct times t_i, where $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_K]$ is the $p \times K$ steering matrix of K linearly independent p-dimensional vectors. The $K \times 1$ vector $\mathbf{s}(t) = [s_1(t), \ldots, s_K(t)]^T$ represents the random signals, assumed zero mean and stationary with full rank covariance matrix. σ is the unknown noise level, and $\mathbf{n}(t)$ is a $p \times 1$ additive Gaussian noise vector, distributed $\mathcal{N}(0, \mathbf{I}_p)$ and independent of $\mathbf{s}(t)$.

Under these assumptions, the population covariance matrix Σ of $\mathbf{x}(t)$ has a diagonal form,

$$\mathbf{W}^H \Sigma \mathbf{W} = \sigma^2 \mathbf{I}_p + \text{diag}(\lambda_1, \ldots, \lambda_K, 0, \ldots, 0) \tag{4.18}$$

where columns of $\mathbf{W}$ forms a basis of $\mathbb{C}^p$ (or of $\mathbb{R}^p$ if the signals are real valued). Let $\mathbf{S}_n$ be the sample covariance matrix of $\{\mathbf{x}_i\}_{i=1}^n$, defined as

$$\mathbf{S}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^H$$

having the eigenvalues $l_1 \geqslant l_2 \geqslant \cdots \geqslant l_p$.

The number of signals K, can then be estimated with the number of eigenvalues of the sample covariance matrix $\mathbf{S}_n$ which are *significanly larger* i.e. bigger than a certain threshold, where the individual thresholds for the eigenvalues can be determined using the

Tracy Widom laws (Theorem 6). The following algorithm, which is deeply motivated by Roy's largest root test (Roy (1953), Johnstone and Nadler (2017)), takes the eigenvalues $l_1, \cdots, l_p$ of the sample covariance matrix $\mathbf{S}_n$ as input and gives the estimated number of signals $\hat{K}_{\text{RMT}}$ as output. The algorithm works as follows: For $k = 1, \ldots, \min(p, n) - 1$, we test

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

Under the null hypothesis, ℓ_k arises from noise. Thus, we reject H_0 if ℓ_k is too large, i.e.

$$\ell_k > \hat{\sigma}^2(k) C_{n,p,k}(\alpha)$$

where $\hat{\sigma}^2(k)$ is an estimate for the unknown noise level σ^2 taken to be,

$$\hat{\sigma}^2(k) = \frac{1}{p-k} \sum_{j=k+1}^{p} l_j \tag{4.19}$$

and

$$C_{n,p,k}(\alpha) = \mu_{n,p-k} + s(\alpha)\xi_{n,p-k} \tag{4.20}$$

where $\mu_{n,p}$ and $\xi_{n,p}$ are the centering and scaling parameters defined as

$$\begin{aligned} \mu_{n,p} &= \frac{1}{n}(\sqrt{n-1/2} + \sqrt{p-1/2})^2 \\ \xi_{n,p} &= \sqrt{\frac{\mu_{n,p}}{n}} \left(\frac{1}{\sqrt{n-1/2}} + \frac{1}{\sqrt{p-1/2}} \right)^{1/3} \end{aligned} \tag{4.21}$$

and $s(\alpha)$ is the $1-\alpha$ quantile of the Tracy Widom distribution. Kritchman and Nadler (2009) showed,

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

Hence, α controls the probability of model overestimation. We stop at the smallest index k where the above condition fails, i.e., the first time we accept H_0. Our estimate of the number of signals is then $\hat{K}_{\text{RMT}} = k - 1$. Hence, the estimator of the number of signals is,

$$\hat{K}_{\text{RMT}} = \arg\min_k \left\{ \ell_k < \hat{\sigma}^2(k)(\mu_{n,p-k} + s(\alpha)\xi_{n,p-k}) \right\} - 1.$$

Algorithm 1: Algorithm for detecting number of signals (Kritchman and Nadler (2009))

Input: Confidence level α, observations ℓ_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 $\hat{\sigma}^2(k) C_{n,p,k}(\alpha)$ using 4.19, 4.20;
  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 $\hat{K}_{\text{RMT}} = \arg\min_k \{\ell_k < \hat{\sigma}^2(k)(\mu_{n,p-k} + s(\alpha)\xi_{n,p-k})\} - 1$;

For a suitably chosen sequence of $\{\alpha\}_n$, $\hat{K}_{RMT,n}$ can be shown to be consistent i.e. $\lim_{n\to\infty} \mathbb{P}(\hat{K}_{RMT,n} = K) = 1$ (Kritchman and Nadler (2009)) where K is the original number of signals.

To demonstrate the performance of the above algorithm, We plot the number of estimated signals when the actual number of signals is in a range of 2 to 5 and the errors are from Standard Normal, $t-$distribution with 5 df, Cauchy and Laplace distribution. Also, we vary the sample size n in a range up to 5000. From Figure 4 it can be seen that, except when the noise has standard Cauchy distribution, if the sample size exceeds 1000, the estimated number of signals is the same as the original number of signals. The algorithm overestimates the number of signals for noise arriving from the Cauchy distribution.

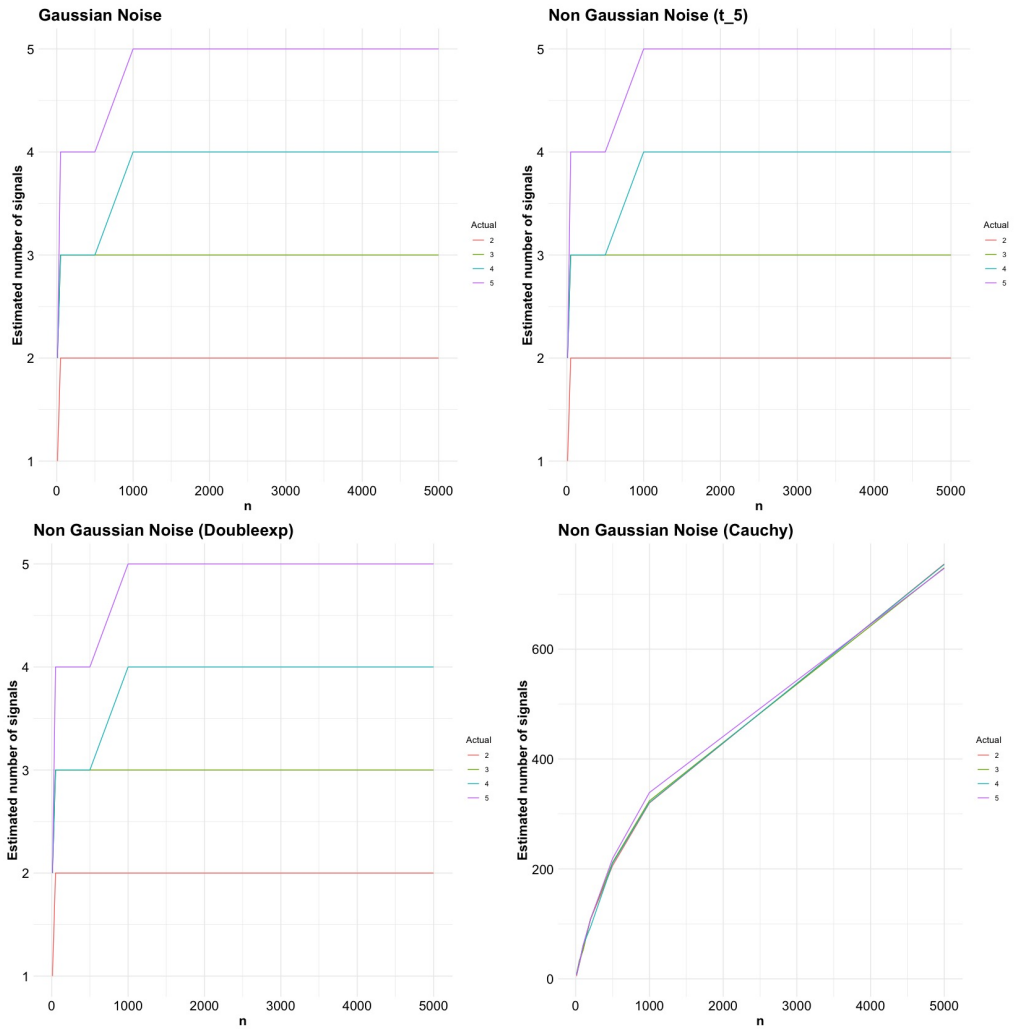


Figure 4: Number of Estimated Signals for different noise distributions: (in clockwise order) Standard Normal, t distribution with 5 df, Standard Cauchy, Double Exponential. The x-axis represents the sample size.

4.4 On Changepoint Detection

Change point detection (CPD) is a statistical method used to identify points in a dataset where the distribution of the data changes significantly. Studies of change-point detection problems date back to 1950. Since then, this topic has been of interest to statisticians and researchers in many other fields such as engineering, economics, climatology, biosciences, genomics, and linguistics due to its diverse applications. Different methods in parametric

and nonparametric setups have been discovered (Niu et al. (2016), Aminikhanghahi and Cook (2017)) for univariate and multivariate time series data. However, when the data is ultrahigh dimensional, most of these traditional methods struggle due to computational complexity or the failure to meet the underlying distributional assumptions. For example, to determine the change of covariance in high dimensional time series data, the sample covariance matrices are used which are extremely large dimensional. In this section, we discuss some methods of detecting change in covariance pattern of high dimensional time series which are motivated by the results of random matrix theory. Change point detection algorithms are traditionally classified as "online" or "offline." We focus on the Offline setting, which considers the entire data set at once and looks back in time to recognize where the change occurred.

The literature on detecting changes in covariance for high dimensional time series has grown substantially in the last few years. Based on the theory of large-scale random matrices in Hero and Rajaratnam (2012), Banerjee et al. (2015) has developed a method for covariance CPD when the data points are independently drawn from an unknown elliptically contoured distribution. Avanesov and Buzun (2018), Wang et al. (2017) obtain method based on the distance between sample covariance matrices, using the operator norm and l_∞ norm of matrices, respectively.

In particular, many authors consider changes in the moderate dimensional setting, that is, where the number of the parameters of the model is of the order of the number of data points. Ryan and Killick (2023) proposes a novel method for detecting changes in the covariance structure of moderate dimensional time series. Let $X_1, \ldots, X_n \in \mathbb{R}^p$ be independent p-dimensional vectors with

$$\operatorname{cov}(X_i) = \Sigma_{i,p}, \quad \text{for } 1 \leq i \leq n,$$

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 $\mathbf{X}_{n,p} := (X_1^T, \ldots, X_n^T)^T$. The method primarily aims to develop a testing procedure that

can identify a change in the covariance structure of the data over time. For now, let us consider the case of a single changepoint. We compare a null hypothesis of the data sharing the same covariance versus an alternative setting that allows a single change at time τ. Formally we have

$$\begin{aligned} H_0 &: \Sigma_{1,p} = \cdots = \Sigma_{n,p} \\ H_1 &: \Sigma_{1,p} = \cdots = \Sigma_{\tau,p} \neq \Sigma_{\tau+1,p} = \ldots \Sigma_{n,p} \end{aligned} \tag{4.22}$$

where τ is unknown. We are interested in distinguishing between the null and alternative hypothesis, and under the alternative locating the changepoint τ, when the dimension of the data p, is of comparable to the sample size, n. In particular, we require that for all pairs n, p, the set

$$T_{n,p}(\ell) := \{t \in \mathbb{Z}^+ \text{ such that } \ell < t < n - \ell\} \quad (2.4)$$

is nonempty, where $\ell > p$ is a problem dependent positive constant. Note $T_{n,p}(\ell)$ defines the set of possible candidate changepoints, while ℓ is the minimum distance between changepoints or minimum segment length. Then for each candidate changepoint $t \in T_{n,p}(\ell)$, a two-sample test statistic $T(t)$ can be used to determine if the data to the left and right of the changepoint have different distributions. If the two sample test statistic for a candidate exceeds some threshold, then we say a change has occurred and an estimator for τ is given by the value $t \in T_{n,p}(\ell)$ that maximizes $T(t)$.

In their method Ryan and Killick (2023), constructs the two sample test statistics using the eigenvalues of the sample covariance matrices of the two samples as follows. For two sample covariance matrices $\mathbf{A}, \mathbf{B}$, in this context, we need to test whether $\mathbf{A}$ and $\mathbf{B}$ are equal or not. So in case they are identical, all of the eigenvalues of $R(\mathbf{A}, \mathbf{B}) := \mathbf{B}^{-1}\mathbf{A}$ is 1. Therefore, the following function of the ratio matrix (or $F-$type matrix) $R(\mathbf{A}, \mathbf{B})$, gives a suitable measure of deviance from the equality of the two matrices,

$$T(\mathbf{A}, \mathbf{B}) = \sum_{j=1}^{p} (1 - \lambda_j(R(\mathbf{A}, \mathbf{B})))^2 + \left(1 - \lambda_j^{-1}(R(\mathbf{A}, \mathbf{B}))\right)^2 \tag{4.23}$$

where $\lambda_j(R(\mathbf{A}, \mathbf{B}))$ is the jth largest eigenvalue of the matrix $R(\mathbf{A}, \mathbf{B})$. The function T has valuable properties that may not be immediately obvious.

Proposition 1 (Ryan and Killick (2023)). *Let $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2 \in \mathbb{R}^{p\times p}$ be the covariance matrices of data $\mathbf{Z}_1 \in \mathbb{R}^{n_1\times p}$ and $\mathbf{Z}_2 \in \mathbb{R}^{n_2\times p}$, respectively, and define T as in (2.5). Then we have that, for any covariance matrix $\mathbf{\Sigma}_0$:*

1. *T is symmetric, that is, $T(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) = T(\mathbf{\Sigma}_2, \mathbf{\Sigma}_1)$;*

2. *T is symmetric with respect to the inversion of matrices, that is,*

$$T(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) = T(\mathbf{\Sigma}_1^{-1}, \mathbf{\Sigma}_2^{-1});$$

3. *If $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_0\mathbf{Z}_1^T\mathbf{Z}_1\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_2 = \mathbf{\Sigma}_0\mathbf{Z}_2^T\mathbf{Z}_2\mathbf{\Sigma}_0$, then*

$$T(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) = T(\mathbf{Z}_1^T\mathbf{Z}_1, \mathbf{Z}_2^T\mathbf{Z}_2).$$

The symmetry property is important for a changepoint analysis as the segmentation should be the same regardless of whether the data is read forward or backward. The second property states that T is the same whether we examine the covariance matrix or the precision matrix. This ensures that differences between both small and large eigenvalues can be detected. The third property is particularly important as we can translate Proposition 1 from two separate datasets $\mathbf{Z}_1, \mathbf{Z}_2$ to two subsets of a single dataset $\mathbf{X}_{n,p}$. This implies that T provides a test statistic that is independent of the underlying covariance of the data. it is to be noted that the function involves ratio matrices which are widely used in multivariate analysis to compare covariance matrices (Finn (1974)). In particular functions of the eigenvalues of the ratio matrices are standard in literature (Wilks (1932), Lawley (1938), Potthoff and Roy (1964)) for inference and methodologies involving covariance matrices. Theorem 7 also discusses the the LSD of the ratio matrics $R(\mathbf{A}, \mathbf{B})$ under suitable conditions. Using

the LSD of the ratio matrices, Ryan and Killick (2023) finds the asymptotic distribution of $T(\mathbf{A}, \mathbf{B})$ for two sample covariance matrices as presented in the following theorem, which gives the framework of the changepoint detection method.

Theorem 12 (Ryan and Killick (2023)). *Let $X_{n_1,p} \in \mathbb{R}^{n_1 \times p}$ and $X_{n_2,p} \in \mathbb{R}^{n_2 \times p}$ be random matrices with independent not necessarily identically distributed entries $\{X_{n_1,i,j}, 1 \leq i \leq n_1, 1 \leq j \leq p\}$ and $\{X_{n_2,k,j}, 1 \leq k \leq n_2, 1 \leq j \leq p\}$ with mean 0, variance 1 and fourth moment $1+\kappa$. Furthermore, for any fixed $\eta > 0$,*

$$\frac{1}{n_1 p} \sum_{j=1}^{p} \sum_{i=1}^{n_1} \mathbb{E}|X_{n_1,i,j}|^4 \mathbf{1}(|X_{n_1,j,k}| \geq \eta\sqrt{n_1}) \to 0 \tag{3.7}$$

$$\frac{1}{n_2 p} \sum_{j=1}^{p} \sum_{i=1}^{n_2} \mathbb{E}|X_{n_2,i,j}|^4 \mathbf{1}(|X_{n_2,j,k}| \geq \eta\sqrt{n_2}) \to 0 \tag{3.8}$$

as n_1, n_2, p tend to infinity such that $\frac{p}{n_1} \to \gamma_1 \in (0,1)$, $\frac{p}{n_2} \to \gamma_2 \in (0,1)$, $\gamma = (\gamma_1, \gamma_2)$ and $\mathbf{1}(\cdot)$ denotes the indicator function. Then as $n \to \infty$,

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

where

$$T(A,B) = \sum_{j=1}^{p} \left[(1 - \lambda_j(B^{-1}A))^2 + (1 - \lambda_j^{-1}(B^{-1}A))^2\right] \ (\lambda_j \text{ is } j\text{th maximum eigenvalue}), \tag{4.24}$$

$$f^*(x) = (1-x)^2 + (1 - 1/x)^2, \tag{4.25}$$

$$\mu(\gamma) = 2K_{3,1}\left(1 - \gamma_2/h^2\right) + 2K_{2,1}\gamma_2/h + 2K_{3,2}\left(1 - \gamma_1^2/h^2\right) + 2K_{2,2}\gamma_1/h, \tag{4.26}$$

$$\sigma^2(\gamma) = \frac{2(K_{2,1}^2 + K_{3,1}^2 + 2K_{3,2}^2)}{h(h^2-1)} + \frac{(J_1 K_{2,1}/h - J_1 K_{3,1}(h^2+1))}{h^2 + (h^2-1)} \tag{4.27}$$

$$+ \frac{(J_2 K_{2,1} 2h)/(h^2-1)^3 + J_2 K_{3,1}(1-3h^2))}{h(h^2-1)^3)} \tag{4.28}$$

$$K_{2,1} = \frac{2h(1+h^2)}{(1-\gamma_2)^4 - 2h/(1-\gamma_2)^2}, \quad K_{2,2} = \frac{2h(1+h^2)^2}{(1-\gamma_1)^4} - 2h/(1-\gamma_1)^2, \tag{4.29}$$

$$K_{3,1} = \frac{h^2}{(1-\gamma_1)^4}, \quad K_{3,2} = \frac{-2(1-\gamma_2)^2}{(1-\gamma_2)^4}, \quad J_2 = (1-\gamma_2)^4, \quad J_1 = -2(1-\gamma_2)^2, \tag{4.30}$$

$$h = \sqrt{\gamma_1 + \gamma_2 - \gamma_1\gamma_2}, \quad \gamma_1 = p/n_1, \quad \gamma_2 = p/n_2, \tag{4.31}$$

$$F_\gamma(dx) = \frac{1-\gamma_2}{2\pi x(\gamma_1+\gamma_2 x)}\sqrt{(b-x)(x-a)}\mathbf{I}_{[a,b]}(x)dx, \tag{4.32}$$

$$a = \frac{(1-h)^2}{(1-\gamma_2)^2}, \quad b = \frac{(1+h)^2}{(1-\gamma_2)^2}. \tag{4.33}$$

So, using Theorem 12 we can immediately have a normalized version of T, i.e.

$$\tilde{T} = \sigma^{-1}(\gamma)\left(T\left(\frac{1}{n_1}X^T_{n_1,p}X_{n_1,p}, \frac{1}{n_2}X^T_{n_2,p}X_{n_2,p}\right) - p\int f^*(x)dF_\gamma(x) - \mu(\gamma)\right) \tag{4.34}$$

which will be asymptotically standard normal, and hence we can use the quantile of standard normal with multiple testing corrections (Haynes (2013)) to test hypothesis 4.22. So using Theorem 12, given one dataset we can test whether the data has one changepoint or not. For the case of Multiple changepoints, the method is generalized using the classic binary segmentation procedure (Scott and Knott (1974)).

The binary segmentation method extends a single changepoint test as follows. First, the test is run on the whole data. While running on a particular interval of time (s, e), for each timepoint τ in that range (except leaving l many timepoints from both sides of the interval, for efficiency purposes as the testing procedure is asymptotic) the algorithm finds the normalized test statistic $\tilde{T}(\tau)$ (as in Equation (4.34)) by breaking the datapoints into two parts pivoting τ and then finds the maximum value of the test statistic $\tilde{T}(\tau)$ over τ in that interval $(s+l, e-l)$ and check if that exceeds a cutoff ν to guarantee the existence of a changepoint in the interval (s, e). 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. Note the threshold, ν, and the minimum segment length, ℓ, remain the same. Note that several

extensions of the traditional binary segmentation procedure have been proposed in recent years (Olshen et al. (2004); Fryzlewicz (2014)) which may be used to generalize the algorithm of Ryan and Killick (2023). The full proposed procedure is described in algorithm 2.

Algorithm 2: Ratio Binary Segmentation (RatioBinSeg) (Ryan and Killick (2023))

Input: Data matrix X, interval (s, e), set of changepoints C, minimum segment length ℓ, significance level α
Output: Set of changepoints C
Set $\nu = \Phi^{-1}(1 - \frac{\alpha}{n^2})$, where $\Phi(\cdot)$ N(0,1) CDF;
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;

In the algorithm, $\bar{\Sigma}$ is the natural estimate of Σ based on the data in the corresponding time interval.

For the multiple changepoint setting, let $\tau := \{\tau_1, ..., \tau_m\}$ and $\hat{\tau} := \{\hat{\tau}_1, ..., \hat{\tau}_m\}$ 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 $1 \leq j \leq \hat{m}$ and 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.

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

Table for FPR for this method and Wang et al. (2017) for various n and p are in Table 2

Table 2: Comparison of FPR for various n, p

		Assumptions of Ryan and Killick (2023)		Assumptions of Wang et al. (2017)	
p	**n**	**Ratio**	**Wang**	**Ratio**	**Wang**
3	500	0.24	0.63	**0.10**	0.64
3	1000	0.28	0.77	**0.14**	0.80
3	2000	0.31	0.85	**0.16**	0.88
3	5000	0.31	0.90	**0.17**	0.92
10	500	0.16	0.27	0.23	0.39
10	1000	0.13	0.43	0.24	0.62
10	2000	0.10	0.53	0.24	0.75
10	5000	**0.09**	0.62	0.19	0.80
30	2000	**0.02**	0.32	**0.03**	0.58
30	5000	**0.02**	0.31	**0.03**	0.78
100	5000	**0.00**	0.45	**0.00**	0.18

5 Conclusion and Future directions

This article highlights the profound role of random matrix theory (RMT) in addressing challenges arising in high-dimensional statistics. By leveraging the asymptotic spectral properties of large random matrices, particularly covariance matrices, and ratios of covariance matrices, RMT provides a novel theoretical foundation for statistical methods. The exploration of both the bulk spectrum and the extreme eigenvalues underscores the versatility of these tools in understanding high-dimensional data structures.

The applications discussed in this article demonstrate the practical relevance of RMT. From inference on covariance matrices to dimensionality reduction through PCA, noise reduction in signal processing, and changepoint detection, RMT proves to be an indispensable framework for tackling modern statistical problems. The unifying principles of RMT not only enhance the theoretical understanding of high-dimensional phenomena but also drive the development of innovative methodologies in diverse fields.

This work provides an inspection of the bridge between the mathematical elegance of random matrix theory and its impactful applications in statistics, emphasizing the potential

for further exploration and development in this vibrant intersection of disciplines. We discuss some of the future directions of application of RMT, which has a great potential:

- Most of the results in RMT are based on iid observations. Though work has been done for certain covariance patterns as well, however, there is a great potential for extending the current theory on the eigenvalues of Wishart-type matrices, when the columns of the data matrix can be viewed as a realization of a high-dimensional multivariate time series, and that can have a significant impact on econometrics and finance.

- The RMT-based methods can be generalized when there are missing values in the dataset. For example, in high-dimensional spatiotemporal statistics, for each time-point, the spatial data is in the form of a matrix. In most cases, for each time point, there are a couple of missing values in the matrix consisting of the spatial data at that time point. In literature, in case it is assumed that the spatial data arises from a random field. Deb et al. (2017) has a detailed discussion about the spectral analysis of such datasets coming from a random field. However, the asymptotic theory provided there assumes the data dimension to be fixed. So generalization of these results using the asymptotic theories of large random matrices is an open problem yet to be solved.

- A potentially useful avenue for the application of RMT is in numerical optimization algorithms that use gradient-based methods for large dimensional data. While there has been explosive growth in mathematical descriptions in the RMT literature, computational tools have not kept pace with the theoretical developments. Integration of computational tools with tools for the analysis of large dimensional data using RMT principles has the potential to create a new paradigm for statistical practices.

References

Aminikhanghahi, S. and Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2):339–367.

Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics*, 34(1):122–148.

Avanesov, V. and Buzun, N. (2018). Change-point detection in high-dimensional covariance structure.

Bai, Z. and Silverstein, J. W. (2007). On the signal-to-interference ratio of cdma systems in wireless communications.

Bai, Z. and Silverstein, J. W. (2010). *Spectral analysis of large dimensional random matrices*, volume 20. Springer.

Bai, Z. and Zhou, W. (2008). Large sample covariance matrices without independence structures in columns. *Statistica Sinica*, pages 425–442.

Bai, Z. D. and Yin, Y. Q. (1988). Convergence to the semicircle law. *The Annals of Probability*, 16(2):863–875.

Bai, Z.-D. and Yin, Y.-Q. (2008). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. In *Advances In Statistics*, pages 108–127. World Scientific.

Bai, Z. D., Yin, Y. Q., and Krishnaiah, P. R. (1988). On the limiting empirical distribution function of the eigenvalues of a multivariate f matrix. *Theory of Probability & Its Applications*, 32(3):490–500.

Baik, J. and Silverstein, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *Journal of multivariate analysis*, 97(6):1382–1408.

Banerjee, T., Firouzi, H., and Hero, A. O. (2015). Non-parametric quickest change detection for large scale random matrices. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pages 146–150. IEEE.

Bao, Z. (2012). Strong convergence of esd for the generalized sample covariance matrices when p/n→ 0. *Statistics & Probability Letters*, 82(5):894–901.

Billingsley, P. (2013). *Convergence of probability measures*. John Wiley & Sons.

Couillet, R. and Debbah, M. (2011). *Random matrix methods for wireless communications*. Cambridge University Press.

Couillet, R. and Liao, Z. (2022). *Random matrix methods for machine learning*. Cambridge University Press.

Deb, S., Pourahmadi, M., and Wu, W. B. (2017). An asymptotic theory for spectral analysis of random fields.

Finn, J. D. (1974). *A general model for multivariate analysis.* Holt, Rinehart & Winston.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). Tthe generalized factor model: identification and estimation. *U%"" 3&*, 4.

Friesen, O., Löwe, M., and Stolz, M. (2013). Gaussian fluctuations for sample covariance matrices with dependent data. *Journal of Multivariate Analysis*, 114:270–287.

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection.

Gotze, F. and Tikhomirov, A. (2006). Limit theorems for spectra of positive random matrices under dependence. *Journal of Mathematical Sciences*, 133:1257–1276.

Hachem, W., Loubaton, P., Mestre, X., Najim, J., and Vallet, P. (2013). A subspace estimator for fixed rank perturbations of large random matrices. *Journal of Multivariate Analysis*, 114:427–447.

Han, X., Pan, G., and Zhang, B. (2016). The tracy–widom law for the largest eigenvalue of f type matrices.

Hastie, T., Buja, A., and Tibshirani, R. (1995). Penalized discriminant analysis. *The Annals of Statistics*, 23(1):73–102.

Haynes, W. (2013). Bonferroni correction. *Encyclopedia of systems biology*, pages 154–154.

Hero, A. and Rajaratnam, B. (2012). Hub discovery in partial correlation graphs. *IEEE Transactions on Information Theory*, 58(9):6064–6078.

Hofmann-Credner, K. and Stolz, M. (2008). Wigner theorems for random matrices with dependent entries: ensembles associated to symmetric spaces and sample covariance matrices.

Hoyle, D. and Rattray, M. (2003). Limiting form of the sample covariance eigenspectrum in pca and kernel pca. *Advances in Neural Information Processing Systems*, 16.

Hui, J. and Pan, G. (2010). Limiting spectral distribution for large sample covariance matrices with m-dependent elements. *Communications in Statistics—Theory and Methods*, 39(6):935–941.

James, O. and Lee, H.-N. (2014). Concise probability distributions of eigenvalues of real-valued wishart matrices. *arXiv preprint arXiv:1402.6757*.

Johnson, D. E. and Graybill, F. A. (1972). An analysis of a two-way model with interaction and no replication. *Journal of the American Statistical Association*, 67(340):862–868.

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

Johnstone, I. M. (2006). High dimensional statistical inference and random matrices. *arXiv preprint math/0611589*.

Johnstone, I. M. and Lu, A. Y. (2009). Sparse principal components analysis. *arXiv preprint arXiv:0901.4392*.

Johnstone, I. M. and Nadler, B. (2017). Roy's largest root test under rank-one alternatives. *Biometrika*, 104(1):181–193.

Karoui, N. E. (2003). On the largest eigenvalue of wishart matrices with identity covariance when n, p and p/n tend to infinity. *arXiv preprint math/0309355*.

Kritchman, S. and Nadler, B. (2009). Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory. *IEEE Transactions on Signal Processing*, 57(10):3930–3941.

Lawley, D. N. (1938). A generalization of fisher's z test. *Biometrika*, 30(1/2):180–187.

Ma, Z. (2012). Accuracy of the tracy–widom limits for the extreme eigenvalues in white wishart matrices.

Marchenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536.

Mardia, K. V., Kent, J. T., and Taylor, C. C. (2024). *Multivariate analysis*, volume 88. John Wiley & Sons.

Muirhead, R. J. (2009). *Aspects of multivariate statistical theory*. John Wiley & Sons.

Nadakuditi, R. R. and Silverstein, J. W. (2010). Fundamental limit of sample generalized eigenvalue based detection of signals in noise using relatively few signal-bearing and noise-only samples. *IEEE Journal of selected topics in Signal Processing*, 4(3):468–480.

Nadler, B. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach.

Niu, Y. S., Hao, N., and Zhang, H. (2016). Multiple change-point detection: a selective overview. *Statistical Science*, pages 611–623.

Olkin, I. and Rubin, H. (1964). Multivariate beta distributions and independence properties of the wishart distribution. *The Annals of Mathematical Statistics*, pages 261–269.

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5(4):557–572.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642.

Paul, D. and Aue, A. (2014). Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference*, 150:1–29.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572.

Perlman, M. D. (1977). A note on the matrix-variate f distribution. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 290–298.

Pinelis, I. and Molzon, R. (2016). Optimal-order bounds on the rate of convergence to normality in the multivariate delta method.

Potthoff, R. F. and Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 51(3-4):313–326.

Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *The Annals of Mathematical Statistics*, 24(2):220–238.

Ryan, S. and Killick, R. (2023). Detecting changes in covariance via random matrix theory. *Technometrics*, 65(4):480–491.

Scott, A. J. and Knott, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, pages 507–512.

Silverstein, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *Journal of Multivariate Analysis*, 55(2):331–339.

Silverstein, J. W. and Combettes, P. L. (1992). Signal detection via spectral theory of large dimensional random matrices. *IEEE Transactions on Signal Processing*, 40(8):2100–2105.

Soshnikov, A. (2002). A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices. *Journal of Statistical Physics*, 108:1033–1056.

Telatar, E. (1999). Capacity of multi-antenna gaussian channels. *European transactions on telecommunications*, 10(6):585–595.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61(3):611–622.

Tracy, C. A. and Widom, H. (1996). On orthogonal and symplectic matrix ensembles. *Communications in Mathematical Physics*, 177:727–754.

Tulino, A. M., Verdú, S., et al. (2004). Random matrix theory and wireless communications. *Foundations and Trends® in Communications and Information Theory*, 1(1):1–182.

Wachter, K. W. (1980). The limiting empirical measure of multiple discriminant ratios. *The Annals of Statistics*, pages 937–957.

Wang, D., Yu, Y., and Rinaldo, A. (2017). Optimal covariance change point localization in high dimension. *arXiv preprint arXiv:1712.09912*.

Wax, M. and Kailath, T. (1985). Detection of signals by information theoretic criteria. *IEEE Transactions on acoustics, speech, and signal processing*, 33(2):387–392.

Wei, M., Yang, G., and Yang, L. (2016). The limiting spectral distribution for large sample covariance matrices with unbounded m-dependent entries. *Communications in Statistics-Theory and Methods*, 45(22):6651–6662.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, pages 471–494.

Wishart, J. (1928). The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, pages 32–52.

Yao, J. (2012). A note on a marčenko–pastur type theorem for time series. *Statistics & probability letters*, 82(1):22–28.

Yin, Y., Bai, Z., and Krishnaiah, P. (1983). Limiting behavior of the eigenvalues of a multivariate f matrix. *Journal of multivariate analysis*, 13(4):508–516.

Yin, Y., Bai, Z., and Krishnaiah, P. (1984). On limit of the largest eigenvalue of the large dimensional sample covariance matrix.

Yin, Y. Q. (1986). Limiting spectral distribution for a class of random matrices. *Journal of multivariate analysis*, 20(1):50–68.

Yin, Y. Q. and Krishnaiah, P. R. (1983). A limit theorem for the eigenvalues of product of two random matrices. *Journal of Multivariate Analysis*, 13(4):489–507.

Yin, Y.-Q. and Krishnaiah, P. R. (1987). Limit theorems for the eigenvalues of a product of large dimensional random matrices when the underlying distribution is isotropic. *Theory of Probability & Its Applications*, 31(2):342–346.

6 Appendix

Proof of Theorem 1: Given $\mathbf{X}_n \sim W_p(\mathbf{\Sigma}, n)$ 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 $\mathbf{\Sigma}$. For a random variable X, let $F_X(\cdot)$ be its CDF and for two CDFs F and G, define

$$\Delta(F, G) := \sup_{x \in \mathbb{R}} |F(x) - G(x)| \tag{6.1}$$

We first prove a couple of lemmas needed for the proof.

Lemma 1: Let X_n, Y_n, X, Y be real-valued random variables having a joint distribution such that (X_n, X) is independent of (Y_n, Y). Then

$$\Delta(F_{X_n+Y_n}, F_{X+Y}) \leqslant \Delta(F_{X_n}, F_X) + \Delta(F_{Y_n}, F_Y) \tag{6.2}$$

Proof of Lemma 1: Fix $x \in \mathbb{R}$. Then observe that,

$$\begin{aligned}
&|\mathbb{P}(X_n + Y_n \leqslant x) - \mathbb{P}(X + Y \leqslant x)| \\
&= |\mathbb{E}\left(\mathbb{P}(X_n \leqslant x - Y_n) - \mathbb{P}(X \leqslant x - Y)|Y_n, Y\right)| \\
&= |\mathbb{E}(\mathbb{P}(X_n \leqslant x - Y_n) - \mathbb{P}(X_n \leqslant x - Y) + \mathbb{P}(X_n \leqslant x - Y) - \mathbb{P}(X \leqslant x - Y)\big|Y_n, Y\big)| \\
&= |\mathbb{E}(\mathbb{P}(X_n \leqslant x - Y_n) - \mathbb{P}(X_n \leqslant x - Y)\big|Y_n, Y\big)| + \mathbb{E}\big|\mathbb{P}(X_n \leqslant x - Y) - \mathbb{P}(X \leqslant x - Y)\big|Y_n, Y\big) \\
&= I + II \ \ (say)
\end{aligned}$$

Observe that $II \leqslant \Delta(F_{X_n}, F_X)$ and

$$\begin{aligned}
I &= \big|\mathbb{P}(X_n + Y_n \leqslant x) - \mathbb{P}(X_n + Y \leqslant x)\big| \\
&= \big|\mathbb{E}\big(\mathbb{P}(Y_n \leqslant x - X_n) - \mathbb{P}(Y \leqslant x - X_n)\big|X_n\big)\big| \\
&\leqslant \mathbb{E}\big|\mathbb{P}(Y_n \leqslant x - X_n) - \mathbb{P}(Y \leqslant x - X_n)\big|X_n\big) \\
&\leqslant \Delta(F_{Y_n}, F_Y)
\end{aligned}$$

Thus for all $x \in \mathbb{R}$, we have $\left|\mathbb{P}(X_n + Y_n \leqslant x) - \mathbb{P}(X + Y \leqslant x)\right| \leqslant \Delta(F_{X_n}, F_X) + \Delta(F_{Y_n}, F_Y)$. Hence **Lemma 1** follows. ■

Lemma 2: Let $\Phi(\cdot)$ be the N(0,1) CDF and m be a natural number. Then there exists $c > 0$ such that

$$\Delta\left(\Phi\left(\sqrt{\frac{m}{2}}\left(e^{\sqrt{\frac{2}{m}}x} - 1\right)\right), \Phi(x)\right) \leqslant \frac{c}{\sqrt{m}} \tag{6.3}$$

Proof of Lemma 2: Let $U_1, \cdots, U_m \overset{iid}{\sim} N(0,1)$. Observe that, $Y_m := \sqrt{\frac{m}{2}} \log\left(1 + \sqrt{2} \cdot \overline{U}_m\right)$ have cdf $F_{Y_m}(x) := \Phi\left(\sqrt{\frac{m}{2}}\left(e^{\sqrt{\frac{2}{m}}x} - 1\right)\right)$ where $\overline{U}_m := \frac{1}{m}\sum_{i=1}^{m} U_i$. Consider, $f(x) := \log(1 + \sqrt{2} \cdot x)$. Since $f(0) = 0$ and $f'(0) = \sqrt{2}$, by theorem 2.10 of Pinelis and Molzon (2016), there exists $c > 0$ such that

$$\sup_{x \in \mathbb{R}} \left|\mathbb{P}\left(\frac{\sqrt{m} f(\overline{U}_m)}{|f'(0)|} \leqslant x\right) - \Phi(x)\right| \leqslant \frac{c}{\sqrt{m}} \tag{6.4}$$

Since $\frac{\sqrt{m} f(\overline{U}_m)}{|f'(0)|} = Y_m$, we have

$$\Delta(F_{Y_m}, \Phi) \leqslant \frac{c}{\sqrt{m}}$$

Hence **Lemma 2** follows. ■

Lemma 3: [Berry Esseen type Bounds for log of χ^2 random variables] Let $Z_m \sim \chi^2_m$. Then, there exists $c > 0$ such that,

$$\sup_{x \in \mathbb{R}} \left|\mathbb{P}\left(\sqrt{\frac{m}{2}} \log\left(\frac{Z_m}{m}\right) \leqslant x\right) - \Phi(x)\right| \leqslant \frac{c}{\sqrt{m}} \tag{6.5}$$

Proof of Lemma 3: Observe that by triangle inequality, we have

$$\Delta(G, \Phi) \leqslant \Delta(G, F_{Y_m}) + \Delta(F_{Y_m}, \Phi) \tag{6.6}$$

where $G(x) := \mathbb{P}\left(\sqrt{\frac{m}{2}} \log\left(\frac{Z_m}{m}\right) \leqslant x\right)$ and F_{Y_m} is as in **Lemma 2**. By **Lemma 2**, there exists $c_1 > 0$, such that $\Delta(F_{Y_m}, \Phi) \leqslant \frac{c_1}{\sqrt{m}}$. By the Berry–Esseen theorem, there exists $c_2 > 0$, such that for all $x \in \mathbb{R}$,

$$\Phi(x) - \frac{c_2}{\sqrt{m}} \leqslant \mathbb{P}\left(\sqrt{\frac{m}{2}}\left(\frac{Z_m}{m} - 1\right) \leqslant x\right) \leqslant \Phi(x) + \frac{c_2}{\sqrt{m}} \tag{6.7}$$

Now $G(x) = \mathbb{P}\left(\sqrt{\frac{m}{2}}\log\left(\frac{Z_m}{m}\right) \leqslant x\right) = \mathbb{P}\left(\sqrt{\frac{m}{2}}\left(\frac{Z_m}{m} - 1\right) \leqslant \sqrt{\frac{m}{2}}\left(e^{\sqrt{\frac{2}{m}}x} - 1\right)\right)$.

So by (6.7), $|G(x) - F_{Y_m}(x)| \leqslant \frac{c_2}{\sqrt{m}}$ i.e. $\Delta(G, F_{Y_m}) \leqslant \frac{c_2}{\sqrt{m}}$. Thus by (6.7), $\Delta(G, \Phi) \leqslant \frac{c}{\sqrt{m}}$ where $c := c_1 + c_2$ completing the proof of **Lemma 3**. ■

Now we complete the proof of the theorem. Observe that $|\mathbf{X}_n| = |\mathbf{\Sigma}|U_1 \cdots U_p$ where $U_j \sim \chi^2_{m-p+j}, j = 1, \cdots, p$; $U_1, \cdots, U_p$ mutually independent and for a matrix $\mathbf{M}$, $|\mathbf{M}|$ denotes the determinant of $\mathbf{M}$. Thus,

$$
\begin{aligned}
&\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\sqrt{\frac{n}{2p}}\left(\sum_{i=1}^{p}\log\left(\frac{\lambda_i^{(n)}}{\lambda_i}\right) - \sum_{i=1}^{p}\log\left(n-p+i\right)\right) \leqslant x\right) - \Phi(x)\right| \\
&= \sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\sqrt{\frac{n}{2}}\left(\sum_{i=1}^{p}\log\left(\frac{U_i}{n-p+i}\right)\right) \leqslant \sqrt{p}x\right) - \Phi(x)\right| \\
&= \sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\sqrt{\frac{n}{2}}\left(\sum_{i=1}^{p}\log\left(\frac{U_i}{n-p+i}\right)\right) \leqslant x\right) - \Phi\left(\frac{x}{\sqrt{p}}\right)\right| \\
&= \sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\sqrt{\frac{n}{2}}\left(\sum_{i=1}^{p}\log\left(\frac{U_i}{n-p+i}\right)\right) \leqslant x\right) - \mathbb{P}(Y_1 + \cdots + Y_p \leqslant x)\right| \quad , Y_1, \cdots, Y_p \overset{iid}{\sim} N(0,1) \\
&\leqslant \sum_{i=1}^{p}\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\sqrt{\frac{n}{2}}\left(\sum_{i=1}^{p}\log\left(\frac{U_i}{n-p+i}\right)\right) \leqslant x\right) - \Phi(x)\right| \quad \text{(by } \textbf{Lemma 1}\text{)} \\
&\leqslant C\sum_{i=1}^{p}\frac{1}{\sqrt{n-p+i}} \quad \text{(by } \textbf{Lemma 3}\text{)} \\
&= O\left(\frac{p}{\sqrt{n}}\right)
\end{aligned}
$$

Hence **Theorem 1** follows. ■