High-dimensional linear models: a random matrix perspective

Jamshid Namdari[1], Debashis Paul[1,*], Lili Wang[2]
1: *Department of Statistics, University of California, Davis*
2: *School of Statistics and Mathematics, Zhejiang Gongshang University*

Abstract

Professor C.R.Rao's *Linear Statistical Inference* is a classic that has motivated several generations of statisticians in their pursuit of theoretical research. This paper looks into some of the fundamental problems associated with linear models, but in a scenario where the dimensionality of the observations is comparable to the sample size. This perspective, largely driven by contemporary advancements in random matrix theory, brings new insights and results that can be helpful even for solving relatively low-dimensional problems. This overview also brings into focus the fundamental roles played by the eigenvalues of large covariance-type matrices in the theory of high-dimensional multivariate statistics.

1 Introduction

Professor C. R. Rao's seminal contributions to all areas of statistics spanning over seven decades have inspired developments of many modern statistical techniques. Two of his books Rao (1952, 1965) are classics in statistical methods and their applications. Statistical inference techniques in linear models, and more broadly, multivariate statistical analysis from a linear models perspective, have been some of the most influential contributions of Rao. His works on testing hypotheses in multivariate analysis and in linear models (Rao, 1948, 1959), on statistical inference for factor analysis (Rao, 1955), on applications of principal components analysis (Rao, 1964), on estimation of variance components (Rao, 1972), and his 1975 Wald Memorial Lectures on estimation theory for linear models (Rao, 1976) are some of the most well-cited works in the statistical literature. In view of this, in this paper we focus our attention to modern theoretical developments in these areas of multivariate statistics with the additional feature that the dimensions of the observation vectors are not fixed, but are allowed to increases with the sample size.

One of the striking features of much of Rao's work in the context of linear statistical inference is the broad generality of the conclusions, and hence their applicability to a multitude of practical problems. In this regard, modern high-dimensional statistics has evolved in broadly two distinct trajectories. One of these branches has utilized the idea of imposing meaningful and interpretable, though not always testable, structural assumptions on the parameters associated with a statistical model – such as sparsity, or low-rankedness – while the other branch attempted to generalize conclusions derived in classical multivariate analysis typically under the Gaussian distribution framework to broader classes, without imposing too many structural constraints on the parameter. The first branch has seen an explosive growth in the literature, driven largely by concurrent development

*Correpondence author: Email: debpaul@ucdavis.edu

in analytical tools for dealing with large-dimensional convex and non-convex optimization problems, and has contributed immensely to the popularity of modern machine learning techniques. Excellent overviews of these developments can be found in the monographs Bühlmann and van de Geer (2011), Hastie, Tibshirani and Wainwright (2015) and Wainwright (2019), among others. Theoretical and methodological developments in the other branch has relied more on the geometric aspects of the data cloud in terms of developing mathematical intuitions and techniques, especially on probabilistic concentration phenomena associated with low-dimensional smooth functions of large-dimensional random vectors, and much less on restrictive model assumptions. However, these developments, greatly influenced by developments in random matrix theory, consequently have been mostly been restricted to settings where the dimensionality of the observations are either comparable or of smaller order than the sample sizes. This is unlike in the first branch where the structural assumptions, such as sparsity, serve to effectively reduce the complexity of the estimation and inference problems, so that nominally the dimensionality of the observations can be orders of magnitude larger than sample size, and yet valid inference is possible. In view of the divergent nature of the current state of high-dimensional statistics, and in order to keep the presentation coherent and focused, in this paper, we present a review of modern multivariate analysis with an emphasis on linear models and related statistical problems, from the perspective of random matrix theory.

Classical multivariate analysis texts, most notably Anderson (2003), Mardia, Kent and Bibby (1980) and Muirhead (1982) have emphasized the use of analysis of fixed dimensional random matrices in analyzing the behavior of various statistics that naturally arise in inference problems. A majority of these problems are naturally formulated in terms of the eigen-decomposition of certain symmetric matrices. These problems can be broadly categorized into two groups. The first group involves the eigen-analysis of a single random symmetric matrix, namely, the sample covariance matrix associated with a set of i.i.d. multidimensional observations. This problem is often referred to as a *single Wishart problem*, after the characterization of such matrices as the multivariate generalization of Chi-square random variables due to Wishart (1928). The second group of inference problems involves analyses of statistics that can be characterized in terms of the eigenvalues of the "ratio" of two independent symmetric covariance-type matrices of the same dimension, and is often referred to as a *double Wishart problem*. This problem is also closely related to the generalized eigenvalue problem associated with a pair of Wishart matrices with same the scale matrix. Multivariate analysis methods that fall within the first group include principal component analysis (PCA) and tests for population covariance matrices in an one-sample problems. Methods falling under the second group include multivariate analysis of variance (MANOVA), canonical correlation analysis (CCA), tests for equality of covariance matrices, and tests for linear hypotheses in multivariate linear regression. In addition, random matrices play a natural role in defining and characterizing estimates in multivariate linear regression problems and in classification and clustering problems. Clearly, most of these problems, including classification (or discriminant analysis) problems under the idealized Gaussian populations framework, are directly or indirectly related to the broader statistical methodology associated with linear models. In view of this, in this paper we give an overview of the theoretical developments dealing with various questions within the context of high-dimensional linear models through the analytical framework of *random matrix theory (RMT)* that is particularly well-suited for the "large dimension and comparable sample size" regime.

In this paper, we spend most of our efforts into dealing with a subset of the topics, namely high-dimensional linear regression (Section 4), MANOVA (Section 5), hypothesis tests on covariances

(Section 6), discriminant analysis (Section 7) and linear time series (Section 8). We also leave out a couple of important topics. One of them pertains to high-dimensional principal components analysis, mainly because there are comprehensive reviews on this topic from the RMT perspective (see for example, (Johnstone and Paul, 2018) and the references therein). The other topic is covariance estimation, and especially estimation of the spectrum of the population covariance matrix within the RMT framework. The latter topic, though important in its own right, does not fit in very well with the rest of the topics, and is hence omitted from the discussion here. Interested readers may look into the following papers to familiarize themselves with some of the distinct approaches to this problem: Donoho, Gavish and Johnstone (2018), El Karoui (2008), Bai, Chen and Yao (2010), Ledoit and Wolf (2012) and Ledoit and Wolf (2015).

2 Ingredients of theoretical analysis

In this section, we summarize some of the fundamental theoretical constructs and objects of study in RMT and a few main technical tools needed to analyze their theoretical properties. In the process, we also give a brief overview of some fundamental results about the behavior of large symmetric random matrices. The reader may refer to the monograph Bai and Silverstein (2010) or the review paper Paul and Aue (2014) for further details.

2.1 Empirical Spectral Distribution (ESD)

Suppose that $\mathbf{X}$ is an $N \times N$ symmetric matrix so that all the eigenvalues $\lambda_1, \ldots, \lambda_N$ of $\mathbf{X}$ are real. The empirical distribution of the eigenvalues of $\mathbf{X}$, usually referred to as the *Empirical Spectral Distribution* (ESD) of $\mathbf{X}$, is the distribution function $N^{-1}\sum_{i=1}^{N}\delta_{\lambda_i}$ where δ_y denotes the Dirac mass at y. We can thus define the empirical spectral (cumulative) distribution function of $\mathbf{X}$ as $F^{\mathbf{X}}(x) = N^{-1}\sum_{j=1}^{N}\mathbf{1}_{\{\lambda_j \le x\}}$ for $x \in \mathbb{R}$. In RMT, the ESD of a random matrix plays a central role in studying the properties of the spectrum. One motivation is that many statistics associated with a random matrix $\mathbf{X}$ can be expressed as a linear functional of its ESD, or a *linear spectral statistic* (LSD), i.e., a function of the form $\int g(x)dF^{\mathbf{X}}(x)$ for some suitably regular function g. Many classical hypothesis testing procedures in multivariate statistics use traces of polynomials or logarithms of the determinants of sample covariance-type matrices, or Fisher-type matrices (to be defined below). These statistics thus can be recognized as LSS associated with appropriate random matrices, which facilitates their intricate theoretical analysis in the large dimensional regimes.

An important question about the ESD that is studied in the RMT literature is whether, after appropriate normalization of the random matrix, this random distribution converges to any specific probability distribution (in an appropriate sense) as the dimension of the matrix grows. The celebrated *semicircle law* provides a first answer to this in the context of *Wigner matrices*, i.e., square Hermitian or real symmetric matrices with independent entries on and above the diagonal with zero mean and unit variance. Wigner (1958) showed that the expected ESD of an $n \times n$ Wigner matrix with Gaussian entries, multiplied by $1/\sqrt{n}$, converges in distribution to the semicircle law that has the p.d.f.

$$f(x) = \frac{1}{2\pi}\sqrt{4 - x^2}\mathbf{1}_{[-2,2]}(x). \tag{1}$$

An analogous result was derived for the sample covariance matrix by Marčenko and Pastur Marčenko and Pastur (1967) assuming that the fourth moments of the entries of the data matrix are finite. The important facet of this result is the dependence of the limiting distribution on the

limiting ratio $c = \lim_{p,n\to\infty} p/n$. Theorem 1 below states Marčenko-Pastur limit law under minimal moment conditions.

We first introduce some basic notations. Henceforth, we use $\mathbf{X}^*$ and and $\mathbf{X}^T$ to denote the conjugate transpose (resp., transpose) for a complex-valued (resp., real-valued) matrix $\mathbf{X}$. Occasionally, for brevity, we use $\mathbf{X}^*$ to mean transpose even when the matrix is real-valued. Also, we use $\mathbf{i}$ to denote the complex number $\sqrt{-1}$, I_N to denote the $N \times N$ identity matrix, and $\Re(z)$ and $\Im(z)$ to denote the real and imaginary parts of a complex number z.

Theorem 1. *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 also that $p, n \to \infty$ such that $p/n \to c \in (0, \infty)$. Then, with probability one, the ESD of $\mathbf{S} = n^{-1}\mathbf{X}\mathbf{X}^*$ converges weakly to a nonrandom distribution, known as the Marčenko-Pastur law, denoted by F_c. If $c \in (0, 1]$, then F_c has a p.d.f.*

$$f_c(x) = \frac{\sqrt{(b_+(c) - x)(x - b_-(c))}}{2\pi c x} \mathbf{1}_{[b_-(c), b_+(c)]}(x), \tag{2}$$

where $b_\pm(c) = (1 \pm \sqrt{c})^2$. If $c \in (1, \infty)$, then F_c is a mixture of a point mass 0 and the p.d.f. $f_{1/c}$ with weights $1 - 1/c$ and $1/c$, respectively.

Theorem 1 is a formal way of quantifying the spreading of the eigenvalues of the sample covariance matrix around their population counterpart. Notice that this spreading of eigenvalues increases as the limiting dimension-to-sample-size ratio c increases from 0 to 1. When $p/n \to 0$, both the largest and the smallest eigenvalue of $\mathbf{S}$ converge to 1 and thus the Marčenko-Pastur law does not hold for the ESD of $\mathbf{S}$.

2.2 Stieltjes transform

The Stieltjes transform plays a central role in RMT in terms of characterizing the asymptotic behavior of eigenvalues of a symmetric random matrix. The Stieltjes transform of a measure μ on the real line is defined as the function

$$\mathcal{S}_\mu(z) = \int \frac{1}{x - z} \mu(dx), \qquad z \in \mathbb{C}^+. \tag{3}$$

where $\mathbb{C}^+ := \{x + \mathbf{i}y \colon x \in \mathbb{R}, y > 0\}$. Note that $\mathcal{S}_\mu$ is analytic on $\mathbb{C}^+$ and maps into $\mathbb{C}^+$. The following inversion formula allows one to reconstruct the distribution function from its Stieltjes transform.

Lemma 1. *Let P be a probability measure on the real line. If $a < b$ are points of continuity of the associated distribution function, then*

$$P((a, b)) = \frac{1}{\pi} \lim_{v \to 0+} \int_a^b \Im(\mathcal{S}_P(u + \mathbf{i}v)) du. \tag{4}$$

The following lemma (Geronimo and Hill, 2003) gives a necessary and sufficient condition for the limit of Stieltjes transforms of a sequence of probability measures to be the Stieltjes transform of a probability measure.

Lemma 2. *Suppose that $\{P_n\}$ is a sequence of Borel probability measures on the real line with Stieltjes transforms $\{s_n\}$. If $\lim_{n\to\infty} s_n(z) = s(z)$ for all $z \in \mathbb{C}^+$, then there exists a Borel probability measure P with Stieltjes transform $\mathcal{S}_P = s$ if and only if*

$$\lim_{v \to \infty} \mathbf{i}v s(\mathbf{i}v) = -1, \tag{5}$$

in which case P_n converges to P in distribution.

To see more clearly the usefulness of Stieltjes transform in the study of the asymptotic behavior of ESDs, suppose that for each $N \geq 1$, $\mathbf{W}_N$ is an $N \times N$ Hermitian random matrix, so that its eigenvalues are all real, with ESD $F^{\mathbf{W}_N}$. Then, the Stieltjes transform of $F^{\mathbf{W}_N}$, say s_N, is given by $s_N(z) = N^{-1}\mathrm{tr}((\mathbf{W}_N - z\mathbf{I}_N)^{-1})$. Notice that $(\mathbf{W}_N - z\mathbf{I}_N)^{-1}$ is the resolvent of the matrix $\mathbf{W}_N$ and its points of singularity are at the eigenvalues of $\mathbf{W}_N$. In view of Lemma 2, in order to prove that the sequence of ESDs $F^{\mathbf{W}_N}$ converges to a distribution F, say (in probability or almost surely), one needs to check that $\{s_N\}$ satisfies the conditions of the lemma (in probability or almost surely).

For problems involving random matrices that are a Wigner or Wishart-type matrix, it is relatively easy to derive an approximate iterative equation for the Stieltjes transform of its ESD. This is typically by using a rank one matrix perturbation technique and an appropriate inversion formulas for block matrices. The reader may see, for example, Sec. 3.1 of Paul and Aue (2014) for a brief overview of this approach.

The following generalization of Theorem 1 due to Silverstein and Bai (1995), especially for the case of real-valued random variables, is of fundamental importance in high-dimensional multivariate analysis.

Theorem 2. *Suppose that the entries of the $p \times n$ matrix $\mathbf{X}_n$ are complex random variables that are independent for each n and identically distributed for all n and satisfy $\mathbb{E}[|X_{11} - E(X_{11})|]^2 = 1$. Also, assume that $\mathbf{A}_p = diag(\tau_1, \ldots, \tau_p)$, where $\tau_j \in \mathbb{R}$ and the empirical distribution function of $\{\tau_1, \ldots, \tau_p\}$ converges almost surely to a probability distribution function H as $n \to \infty$. Let $\widetilde{\mathbf{W}}_n = \mathbf{B}_n + n^{-1}\mathbf{X}_n^*\mathbf{A}_p\mathbf{X}_n$ where $\mathbf{B}_n$ is an $n \times n$ Hermitian matrix satisfying that $F^{\mathbf{B}_n}$ converges to $F^{\mathbf{B}}$ almost surely, where $F^{\mathbf{B}}$ is a distribution function on $\mathbb{R}$. Assume further that $\mathbf{X}_n$, $\mathbf{A}_p$ and $\mathbf{B}_n$ are independent. Then as $n, p \to \infty$ such that $p/n \to c \in (0, \infty)$, the ESD $F^{\widetilde{\mathbf{W}}_n}$ of $\widetilde{\mathbf{W}}_n$ converges to a nonrandom distribution $\widetilde{F}_c$, where, for any $z \in \mathbb{C}^+$, its Stieltjes transform $\underline{s} = \underline{s}(z)$ is the unique solution in $\mathbb{C}^+$ of the equation*

$$\underline{s} = s_B\left(z - c\int \frac{\tau dH(\tau)}{1 + \underline{s}\tau}\right), \tag{6}$$

where $s_B(z)$ is the Stieltjes transform of $F^{\mathbf{B}}$.

When $\mathbf{B}_n$ is the zero matrix, (6) reduces to

$$z = -\frac{1}{\underline{s}} + c\int \frac{\tau dH(\tau)}{1 + \underline{s}\tau} \tag{7}$$

which gives an explicit inverse function for $\underline{s}(z)$. Assume further that $\mathbf{A}_n$ is positive definite. Defining $s(z) = (1/c)(\underline{s}(z) + (1-c)/z)$ and noticing that the nonzero eigenvalues of

$$\mathbf{W}_n := \frac{1}{n}\mathbf{A}_p^{1/2}\mathbf{X}_n\mathbf{X}_n^*\mathbf{A}_p^{1/2} \tag{8}$$

coincide with those of $n^{-1}\mathbf{X}_n^*\mathbf{A}_p\mathbf{X}_n$, it can be easily deduced that, under the assumptions of Theorem 2, the ESD $F^{\mathbf{W}_n}$ of $\mathbf{W}_n$ converges to a nonrandom distribution F_c almost surely, where the Stieltjes transform $s = s(z)$ of F_c is the unique solution in the set $\{s \in \mathbb{C}\colon -(1-c)/z + cs \in \mathbb{C}^+\}$ of

$$s = \int \frac{dH(\tau)}{\tau(1 - c - czs) - z}\ . \tag{9}$$

Equations (7) and (9) are variously referred to as the *Marčenko-Pastur equations* or *Silverstein equations*. These form the building blocks for downstream analyses about the analytic behavior of the *Limiting Spectral Distribution (LSD)* of covariance matrices. Notably, Silverstein and Choi (1995) characterized the analytic properties of the limiting distribution F_c, in particular, proving the existence of a continuous density on $\mathbb{R}^+$ for F_c when $c \in (0,1)$, and determining the support of F_c in terms of the zeros of the derivative, with respect to s (restricted to $\mathbb{R}$), of z satisfying (7).

2.3 Linear Spectral Statistics (LSS)

Let F_n be the ESD of a $p \times p$ symmetric or Hermitian random matrix with eigenvalues $\lambda_1, \ldots, \lambda_p$. We define a *linear spectral statistics (LSS)* corresponding to the function g defined on the real line to be the quantity

$$\int g(x) dF_n(x) = \frac{1}{p} \sum_{j=1}^{p} g(\lambda_j). \tag{10}$$

If, as $n \to \infty$, $p \to \infty$ and F_n converges to a limiting distribution F (in probability or almost surely) and g is a continuous function, then $\int g(x) dF_n(x) \to \int g(x) dF(x)$ (in probability or almost surely, respectively). Moreover, it is of interest to study the fluctuations of $\int g(x) dF_n(x)$ around $\int g(x) dF(x)$. The process $G_n(x) = \alpha_n(F_n(x) - F(x))$ can be viewed as a random element in $\mathbb{D}[0,\infty)$ (the metric space of functions with discontinuities of the first kind along with the Skorohod metric), for an appropriate normalizing sequence $\alpha_n \to \infty$. However, this process in general does not converge to a limiting process (Bai and Silverstein, 2004).

However, it may still be possible to find an appropriate sequence α_n such that the random variables

$$G_n(g) := \alpha_n \int g(x)(dF_n(x) - dF(x)) \tag{11}$$

may converge to some limit law for a suitably regular class of functions g. If F_n is the ESD of a Wishart$(p, n; I_p)$ matrix and $p, n \to \infty$ such that $p/n \to c \in (0,1)$, this is indeed the case for "nice" functions such as $g(x) = x^r$ for a positive integer r. The result holds in much greater generality.

Bai and Silverstein (2004) proved a CLT for the random vector $(G_n(g_1), \ldots, G_n(g_K))$ with $\alpha_n = n$ under the $p/n \to c > 0$ setting, when the following assumptions hold.

(i) $F_n = F^{\mathbf{W}_n}$ is the ESD of $\mathbf{W}_n = n^{-1}\mathbf{A}_p^{1/2}\mathbf{X}_n\mathbf{X}_n^*\mathbf{A}_p^{1/2}$, where $\mathbf{A}_p$ is a $p \times p$ random positive definite matrix whose ESD converges to a nonrandom distribution $F^{\mathbf{A}}$, and is independent of the $p \times n$ matrix $\mathbf{X}_n$ with i.i.d. real- or complex-valued entries X_{jk} satisfying $\mathbb{E}[X_{11}] = 0$, $\mathbb{E}[|X_{11}|^2] = 1$, $\mathbb{E}[|X_{11}|^4] < \infty$; and $\mathbb{E}[X_{11}^4] = 3$ if X_{11} and $\mathbf{A}_p$ are real, while $\mathbb{E}[X_{11}^2] = 0$ and $\mathbb{E}[|X_{11}^4|] = 2$ if X_{11} is complex;

(ii) $g_1, \ldots, g_K$ are analytic functions on an open interval containing $[\liminf \lambda_{\min}^{\mathbf{A}} \mathbf{1}_{(0,1)}(c)(1 - \sqrt{c})^2, \limsup \lambda_{\max}^{\mathbf{A}}(1 + \sqrt{c})^2]$, where $\lambda_{\min}^{\mathbf{A}}$ and $\lambda_{\max}^{\mathbf{A}}$ denote the smallest and the largest eigenvalues of $\mathbf{A}_p$, respectively.

A central idea in the proof is the following representation which uses the *Cauchy integral formula*:

$$G_n(g) = -\frac{1}{2\pi\mathbf{i}} \oint_{\mathcal{C}} g(z)(n(s_n(z) - s(z)))dz,$$

where $\mathcal{C}$ is a positively oriented contour enclosing the support of F and F_n, and s and s_n are the Stieltjes transforms of F and F_n, respectively. So, the problem of finding the limit distribution

of $G_n(g)$ reduces to studying the asymptotic behavior of the process of $M_n(z) = n(s_n(z) - s(z))$. Bai and Silverstein (2004) constructed a truncated version of $M_n(z)$ on a suitably chosen contour, and applied martingale decomposition techniques on the latter process to derive the final result. In contrast to the Stieltjes transform-based approaches, Lytova and Pastur (2009) used the Fourier transform of the LSS as the basic building block. They first proved the results for matrices with Gaussian entries and then used an interpolation between the random matrix ensemble of interest and a conveniently chosen Gaussian matrix, and then used a version of *Stein's Lemma*, for proving the result.

CLTs for certain special classes of linear spectral statistics, for example the log-determinant of the sample covariance matrix, have been proved in many different contexts. Zheng (2012) proved a CLT for the LSS of Fisher-type matrices (or F-type, or double Wishart matices). Specifically, she considered asymptotic behavior of the ESD of matrices of the form $\mathbf{M}_n = (n_1^{-1}\mathbf{X}\mathbf{X}^T)(n_2^{-1}\mathbf{Y}\mathbf{Y}^T)^{-1}$, with $\mathbf{X} = [X_{\cdot 1} : \ldots : X_{\cdot n_1}]$, $\mathbf{Y} = [Y_{\cdot 1} : \ldots : Y_{\cdot n_1}]$ representing independent samples of sizes n_1 and n_2 from two p-dimensional populations with i.i.d. zero mean, unit variance entries. In this case, by making use of the Marčenko-Pastur equation, it is possible to show the existence of an LSD F_{c_1,c_2} for the matrix $\mathbf{M}_n$, assuming that $p/n_1 \to c_1 \in (0,\infty)$, and $p/n_2 \to c_2 \in (0,1)$. For an integrable function g, let $F_{c_{n1},c_{n2}}(g) = \int g(x)dF_{c_{n1},c_{n2}}(x)$, where $c_{n1} = p/n_1$ and $c_{n2} = p/n_2$, and $F_{c_{n1},c_{n2}}$ is the c.d.f. F_{c_1,c_2} with (c_1, c_2) replaced by (c_{n1}, c_{n2}). Under the stated framework, Zheng (2012) derived the joint asymptotic normality for functionals of the type $X_n(g_j) = p(g_j(\mathbf{M}_n) - F_{c_{n1},c_{n2}}(g_j))$, $j = 1, \ldots, J$, where $\{g_j\}_{j=1}^J$ is a collection of functions analytic on an open interval containing the support of F_{c_1,c_2}.

CLTs for linear spectral statistics, involving both Wishart-type and F-type matrices, have been extensively used to study the asymptotic behavior of commonly used statistics such as the likelihood ratio statistic, or statistics based on traces of powers of respective matrices in the "large dimension, comparable sample size" settings. In particular, these analyses demonstrate three different features:

(i) Classical asymptotic approximations of the test statistics based on the "fixed dimension, large sample size" regime are inadequate in capturing the effect of dimensionality in the RMT regimes. Also, the distributional approximations become invalid in such set up.

(ii) The approximations based on the RMT framework (almost exclusively built around either CLT for LSS involving these matrices, or based on approximations to distributions of extreme eigen-values) quite accurately capture the bias in measures of goodness-of-fit arising due to dimensionality over a fairly wide range of dimension-to-sample size ratio.

(iii) Moreover, many of these phenomena depend primarily on the geometry of the data cloud and concentration phenomena associated with quadratic forms of large-dimensional random vectors, and so are somewhat robust to distributional assumptions.

2.4 Extreme eigenvalues and Tracy-Widom laws

Following up on the work of Wigner (1958), in the process of analyzing the behavior of mathematical models for heavy nuclei, Tracy and Widom (Tracy and Widom, 1994a),Tracy and Widom (1994b), Widom (1999)) derived the limiting distributions of the largest eigenvalue of a Wigner matrix, and derived a class of distributions as limit laws that came to be known as the *Tracy-Widom distributions*. Johnstone (2001) showed that the largest eigenvalues for real and complex-valued Wishart matrices, after appropriate normalization, as $p/n \to c \in (0,1)$, also converge to the

corresponding *Tracy-Widom laws*, as $p, n \to \infty$ such that $p/n \to \gamma \in (0, 1]$ (and so for $\gamma \in [1, \infty)$, by interchanging the role of n and p).

Below we summarize the results describing the distributional limits of the extreme eigenvalues of various classical random matrix ensembles. We first give a very brief description of these ensembles in terms of the joint distribution of their eigenvalues.

The Gaussian Orthogonal Ensemble (GOE) (resp., Gaussian Unitary Ensemble (GUE)) in dimension N is defined as a probability density function on the space of $N \times N$ symmetric matrices (resp., Hermitian matrices), given by

$$h_{N,\beta}(H) = K_{N,\beta} e^{-\beta \mathrm{tr}(H^2)/4},$$

where $K_{N,\beta}$ is constant of proportionality, with $\beta = 1$ corresponding to the GOE and $\beta = 2$ corresponds to the GUE. Using standard arguments (Muirhead, 1982), and the orthogonal (resp. unitary) invariance of the corresponding matrix variates, these ensembles, as well as the associated *Laguerre* and *Jacobi* ensembles described below, are also equivalently expressed in terms of the joint densities of their eigenvalues. A general form of the joint density of the eigenvalues $x_1, ..., x_N$ of H is given by

$$f_{N,\beta,w}(x_1, ..., x_N) = c_{N,\beta,w} \prod_{j<k} |x_j - x_k|^\beta \prod_j (w(x_j))^{\beta/2}, \tag{12}$$

where $c_{N,\beta,w}$ is a proportionality constant and $w(x)$ is a nonnegative weight function, and $\beta = 1$ and $\beta = 2$ correspond to the symmetric (orthogonal) and Hermitian (unitary) matrix ensembles, respectively. For GOE (corresponding to $\beta = 1$) and GUE (corresponding to $\beta = 2$), the weight function $w(x) = \exp(-x^2/2)$.

Laguerre Orthogonal Ensemble (LOE) and *Laguerre Unitary Ensemble (LUE)* refer to the joint density of eigenvalues of the (unnormalized) sample covariance matrix associated with an $N \times (N + \alpha)$ data matrix with i.i.d. real or complex standard Gaussian entries. This joint density has the form (12), (with LOE corresponding to $\beta = 1$ and LUE corresponding to $\beta = 2$), where the index α appears in the expression for the corresponding weight function $w(x) = x^\alpha \exp(-x)\mathbf{1}_{(x>0)}$.

With parameters $N = p$, $\alpha = m - p$ and $\gamma = n - p$, and assuming $\min\{m, n\} > p$, the *Jacobi Orthogonal Ensemble (JOE)* and *Jacobi Unitary Ensemble (JUE)* refer to the joint density of eigenvalues of the matrix $\mathbf{T} = \mathbf{U}(\mathbf{U} + \mathbf{V})^{-1}$ where $\mathbf{U} = \mathbf{X}\mathbf{X}^*$ and $\mathbf{V} = \mathbf{Y}\mathbf{Y}^*$, where $\mathbf{X}$ and $\mathbf{Y}$ are independent $p \times m$ and $p \times n$ matrices, with i.i.d. real (corresponding to orthogonal) or complex (corresponding to unitary) standard Gaussian entries. For this reason, these ensembles are also referred to as "double Wishart" ensembles. Notice that the p.d.f. given by (12) for the JOE ($\beta = 1$)/JUE ($\beta = 2$), with the weight function $w(x) = (1 - x)^\alpha (1 + x)^\gamma \mathbf{1}_{(-1<x<1)}$, is actually the joint density of eigenvalues of the transformed matrix $2\mathbf{T} - I_N$ (Johnstone, 2008).

The c.d.f. of the Tracy-Widom distribution associated with the *Gaussian Unitary Ensemble (GUE)* and *Laguerre Unitary Ensemble (LUE)*, denoted by F_2, is given by

$$F_2(s) = \exp\left(-\int_s^\infty (x - s)q^2(x)dx\right), \qquad s \in \mathbb{R}; \tag{13}$$

while the c.d.f. of the Tracy-Widom distribution corresponding to the *Gaussian Orthogonal Ensemble (GOE)* and Laguerre Orthogonal Ensemble (LOE), denoted by F_1, is given by

$$F_1(s) = \exp\left(-\frac{1}{2}\int_s^\infty (q(x) + (x - s)q^2(x))dx\right), \qquad s \in \mathbb{R}; \tag{14}$$

where $q(x)$ satisfies the Painlevé II differential equation $q''(x) = xq(x) + 2q^3(x)$ with the feature that $q(x) - A(x) \to 0$ as $x \to \infty$, where $A(x)$ denotes the Airy function. The main results of Johnstone (2001) can now be stated as follows.

Theorem 3. *Suppose that the entries of the $p \times n$ matrix $\mathbf{X}$ are i.i.d. complex Gaussian with mean zero and variance one. Let $l_{1,p}$ denote the largest eigenvalue of $\mathbf{X}\mathbf{X}^*$. If $p/n \to \gamma \in (0, 1]$, as $n \to \infty$, then*

$$\frac{l_{1,p} - \mu_{n,p}}{\sigma_{n,p}} \implies W_2, \tag{15}$$

where

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

If the the entries of $\mathbf{X}$ are i.i.d. real Gaussian with mean zero and variance one, and $l_{1,p}$ denotes the largest eigenvalue of $\mathbf{X}\mathbf{X}^T$, then as $n \to \infty$, so that $p/n \to (0, 1]$,

$$\frac{l_{1,p} - \mu'_{n,p}}{\sigma'_{n,p}} \implies W_1, \tag{16}$$

where

$$\mu'_{n,p} = (\sqrt{n-1} + \sqrt{p})^2, \qquad \sigma'_{n,p} = (\sqrt{n-1} + \sqrt{p}) \left(\frac{1}{\sqrt{n-1}} + \frac{1}{\sqrt{p}} \right)^{1/3}.$$

Here the random variables W_1 and W_2 are distributed with c.d.f. F_1 and F_2, respectively.

Johnstone (2008) proved similar scaling limits for the largest eigenvalues of *Jacobi Orthogonal Ensemble (JOE)* and *Jacobi Unitary Ensemble (JUE)*. Denote by $\theta_{1,p}$ the largest eigenvalue of $\mathbf{U}(\mathbf{U} + \mathbf{V})^{-1}$, where $\mathbf{U} = \mathbf{X}\mathbf{X}^*$ and $\mathbf{V} = \mathbf{Y}\mathbf{Y}^*$ with $\mathbf{X}$ $(p \times m)$ and $\mathbf{Y}$ $(p \times n)$ being independent matrices with i.i.d. real or complex standard Gaussian entries. Then under the assumption that

$$m = m(p) \to \infty,\ n = n(p) \to \infty \text{ as } p \to \infty \text{ such that } \lim_{p\to\infty} \frac{\min\{p, n\}}{m + n} > 0 \text{ and } \lim_{p\to\infty} \frac{p}{m} < 1, \tag{17}$$

the normalized quantity $[\log(\theta_{1,p}/(1-\theta_{1,p})) - \mu_p]/\sigma_p$, converges in distribution to the Tracy-Widom laws F_2 and F_1, in the complex and real settings, respectively, where μ_p and σ_p are appropriate centering and scaling sequences,

Various generalizations of these results, in particular statements that these results are *universal*, i.e., do not depend heavily on the distribution of the observations, have appeared in the literature, especially due to the work of Terence Tao, Van Vu, Horng-Tzer Yau and coworkers, a brief summary of which can be found in Johnstone and Paul (2018).

3 Inference problems in high-dimensional linear models

A fundamental problem in statistics, not just restricted to linear models, is to compare two or more populations of multivariate observations based on samples drawn from these populations. An idealization of this problem, when the measurements are continuous, is to assume that we have independent observations from $k \geq 2$ p-variate normal distributions such that l-th population has mean $\boldsymbol{\mu}_l$ and covariance matrix Σ_l. If furthermore, these populations potentially differ only due to differences in the locations, while having a common scale, i.e., $\Sigma_1 = \cdots = \Sigma_k$, then the

resulting statistical problem of comparing these k populations becomes a Multivariate Analysis of Variance (MANOVA) problem, where the primary objective is to distinguish among the k different populations based on formulating and testing hypotheses about the means $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$, or some contrasts involving these parameters.

A generalization of this MANOVA problem can be formulated in terms of testing hypotheses about pre-specified linear functionals of the regression parameter $\mathbf{B}$ in a multivariate linear regression model

$$\mathbf{Y} = \mathbf{B}\mathbf{X} + \boldsymbol{\varepsilon} \tag{18}$$

where $\mathbf{Y}$ is a $p \times n$ matrix whose columns consist of measurements on p response variables, $\mathbf{B}$ is a $p \times k$ matrix of regression coefficients, $\mathbf{X}$ is a $k \times n$ *fixed* design matrix, and $\boldsymbol{\varepsilon}$ is a $p \times n$ matrix of noise terms, whose columns are assumed to be of the form $\varepsilon_{\cdot j} = \Gamma Z_{\cdot j}$, where Z_{ij}'s may be taken to be i.i.d. random variables with zero mean and unit variance. Then the covariance of the noise is $\Gamma\Gamma^T = \Sigma$, say. In this framework, the pre-specified linear hypotheses involving the regression parameter $\mathbf{B}$ can be collectively formulated as a single hypothesis of the form $H_0 : \mathbf{BC} = \mathbf{O}$ where $\mathbf{C}$ is a $k \times q$ matrix (may be referred to as a "contrast matrix") and $\mathbf{O}$ is a $p \times q$ matrix of zeros.

Even in the standard set up of univariate linear regression, when the number of predictor variables is large compared to the sample size, some commonly accepted phenomena associated with classical multivariate analysis, such as consistency of the least squares estimator, consistency of inference procedures on the regression coefficient, etc., fail to hold. Such phenomena can also be explained by utilizing the theoretical framework of RMT. At the same time, commonly used regularization procedures for taking into account the effects of large dimensionality, such as ridge regression, can be understood from a more precise geometric perspective, by making use of such an analytical approach.

Another exciting theoretical development in recent times has been the use of RMT framework to understand robust multivariate procedures in high-dimensional settings. These developments merge two different core principles, one based on a representation of solutions of complex optimization procedures, and the other based on the concentration phenomena associated with large-dimensional random vectors, to provide sophisticated, yet more interpretable, description of the behavior of robust estimators of the regression coefficients and their risk properties.

In the following two sections, we shall study various facets of high-dimensional phenomena associated with commonly used multivariate statistical techniques for linear regression and MANOVA problems. We also review some of the solutions proposed to address the limitations of the existing inferential techniques whose validity rests upon classical "fixed dimension, large sample size" asymptotic regime. In the context of the linear regression model (18), which is our first topic of discussion, beyond inference on the regression coefficient, we shall also look into some associated statistical problems, such as measures of fit and effects of ridge-type regularization schemes. We shall address the MANOVA problems in the high-dimensional setting, namely, when p is comparable to n, the total sample size, while the number of different populations can be either fixed, or growing proportionately to the sample size. Apart from describing the dimension-dependent correction procedures needed by some well-known inference techniques, we shall also review some modern developments involving spectral regularization techniques. Some of the materials present here follow the overall descriptions of the problems in Yao, Zheng and Bai (2015). Other resources include the monographs of Bai and Silverstein (2010) and Fujikoshi, Ulyanov and Shimazu (2011).

4 Linear regression

Over last two decades, there has been a very substantial enhancement in our understanding of the univariate and multivariate linear regression problems within the RMT framework of dimensions comparable to sample sizes. There are many facets to this broad framework. To give a glimpse of the multitude of developments we provide representative summaries of four separate problems involving linear regression. They illustrate high-dimensional phenomena associated with some key statistical problems associated with linear models, namely, (i) behavior of quantitative measures of goodness of fit; (ii) inference on the regression coefficient from the point of view of determining adequacy of submodels; (iii) regularization schemes to account for dimensionality effects and their implications to estimation and prediction performance; (iv) behavior of robust estimators of the regression coefficient; and (v) behavior of estimators of variance components in high-dimensional random and mixed effects models.

4.1 Multiple correlation coefficient

Consider a random regression model where we have i.i.d. realizations of p-dimensional random vector X with population covariance matrix Σ, which is assumed to be positive definite. Then the *squared multiple correlation coefficient* ρ^2 between the variable X_1 and all the other variables $(X_2, \ldots, X_p)$ is given by

$$\rho^2 = \frac{\boldsymbol{\sigma}_1^T \Sigma_{22}^{-1} \boldsymbol{\sigma}_1}{\sigma_{11}} = \frac{\boldsymbol{\beta}^T \Sigma_{22} \boldsymbol{\beta}}{\sigma_{11}} \ . \tag{19}$$

Here, $\sigma_{11} = \mathrm{Var}(X_1)$ is the $(1,1)$ element of Σ, Σ_{22} is $(p-1) \times (p-1)$ dimensional covariance matrix of $(X_2, \ldots, X_p)$, $\boldsymbol{\sigma}_1$ is the $(p-1) \times 1$ vector of covariances $\mathrm{Cov}(X_1, X_j)$, $j = 2, \ldots, p$, and $\boldsymbol{\beta} = \Sigma_{22}^{-1} \boldsymbol{\sigma}_1$. If, without loss of generality, we assume $\mathbb{E}(X)$ is zero, then $\boldsymbol{\beta}$ is the population regression coefficient for the regression of X_1 on $X_{-1} := (X_2, \ldots, X_p)^T$, i.e., $\boldsymbol{\beta} = \mathbb{E}(X_1 | X_{-1})$. Note that $\rho^2 \in [0,1]$ and has the interpretation as the proportion of variability in X_1 explained by X_{-1} through a linear regression model.

Now suppose that we have sample observations $X_{\cdot j}$, $j = 1, \ldots, n$ from the p-variate normal distribution $\mathcal{N}_p(\mathbf{0}, \Sigma)$. Then, defining $\widehat{\Sigma} = n^{-1} \sum_{j=1}^n (X_{\cdot j} - \overline{X})(X_{\cdot j} - \overline{X})^T$, where $\overline{X} = n^{-1} \sum_{j=1}^n X_{\cdot j}$ is the sample mean, and correspondingly defining $\widehat{\sigma}_{11}$, $\widehat{\boldsymbol{\sigma}}_1$ and $\widehat{\boldsymbol{\beta}} = \widehat{\Sigma}_{22}^{-1} \widehat{\boldsymbol{\sigma}}_1$ (estimate of $\boldsymbol{\beta}$), the sample squared multiple correlation coefficient between variables X_1 and X_{-1} is defined as

$$R^2 = \frac{\widehat{\boldsymbol{\sigma}}_1^T \widehat{\Sigma}_{22}^{-1} \widehat{\boldsymbol{\sigma}}_1}{\widehat{\sigma}_{11}} = \frac{\widehat{\boldsymbol{\beta}}^T \widehat{\Sigma}_{22} \widehat{\boldsymbol{\beta}}}{\widehat{\sigma}_{11}} \ . \tag{20}$$

Traditionally, R^2 has been used a simple and effective measure of the strength of linear association in linear regression models. A more widely used statistic for comparison across models, that adjusts for the number of variables, is the *adjusted* R^2, defined as $R^2_{adj} = \frac{n-1}{n-p} R^2$. Under Gaussianity of observations, it is well-known that the transformed statistic $\{R^2/(p-1)\}/\{(1-R^2)/(n-p)\}$ has a central F distribution with $(p-1, n-p)$ degrees of freedom if $\rho = 0$ (i.e., there is no linear association between X_1 and the rest of the variables). Indeed, the invariance of this distribution with respect to Σ under the setting $\rho = 0$ is one of the reasons as to why R^2, or some transformed version of it, has traditionally been used for testing the presence of linear association between the response and predictor variables.

In the traditional "fixed p, large n" asymptotic regime, as $n \to \infty$, both R^2 and R^2_{adj} are consistent estimators of ρ^2, which follows trivially from the consistency of $\widehat{\boldsymbol{\Sigma}}$ as an estimator of $\boldsymbol{\Sigma}$.

However, this is not the case when p and n are of comparable sizes. In particular, when $p, n \to \infty$ such that $p/n \to c \in (0,1)$, we have the following result due to Yao, Zheng and Bai (2015) about the pointwise and distributional limits of R^2.

Theorem 4. *Suppose that $X_{\cdot 1}, \ldots, X_{\cdot n} \sim \mathcal{N}_p(\mathbf{0}, \Sigma)$, with a positive definite Σ. Suppose further that $p.n \to \infty$ such that $c_n := p/n \to c \in (0,1)$. Then*

$$R^2 \xrightarrow{a.s.} c + (1-c)\rho^2, \tag{21}$$

and

$$\sqrt{n}(R^2 - c_n - (1-c_n)\rho^2) \xRightarrow{D} \mathcal{N}(0, \sigma^2(c, \rho^2)) \tag{22}$$

where $\sigma^2(c,t) = 2(c+(1-c)t) - 2(2c + 4(1-c)t - 2(1-c)t^2)(c + (1-c)t - 1/2)$.

Notice that (21) quantifies the asymptotic bias in R^2 as an estimator of ρ^2, while (22) can be used to perform tests for the hypothesis $\rho = \rho_0$ for any pre-specified $\rho_0 \in [0,1)$, and such a test can then be inverted to find a confidence interval for ρ.

4.2 Inference in multivariate linear regression

Testing for the presence of submodels of lower complexity is one of the fundamentals problems in linear models. In the context of linear regression, typically we are interested to see if a subset of the predictors is adequate in terms of describing the variability in the response. Bai et al. (2013) considered the problem of testing linear constraints associated with the regression coefficient matrix $\mathbf{B}$ in the high-dimensional multivariate linear regression model (18) under Gaussianity. Specifically, they proposed appropriate corrections to the likelihood ratio test (LRT) for hypotheses of the form $H_0 : \mathbf{BC} = \mathbf{D}_0$, where $\mathbf{C}$ is a matrix of full column rank (constraints matrix), and $\mathbf{D}_0$ denotes a pre-specified matrix. Below, we give a simplified description of this problem and the associated theoretical analysis when the constraint simply sets a subset of the columns of $\mathbf{B}$ to a fixed value.

Suppose we partition the $k \times n$ design matrix $\mathbf{X}$ into two parts, $\mathbf{X}_1$ and $\mathbf{X}_2$, consisting of first k_1 and last $k_2 = (k - k_1)$ rows of $\mathbf{X}$, respectively. Let the corresponding partition of $\mathbf{B}$ be $\mathbf{B} = [\mathbf{B}_1 : \mathbf{B}_2]$ where $\mathbf{B}_l$ is $p \times k_l$, $l = 1, 2$. Suppose we are interested in testing the null hypothesis: $H_0 : \mathbf{B}_1 = \mathbf{B}_{10}$, where the latter is a pre-specified matrix. If $\mathbf{B}_{10}$ is a matrix of zero entries, then the hypothesis testing problem can be seen as a model selection problem, where the submodel $\mathbf{Y} = \mathbf{B}_2\mathbf{X}_2 + \boldsymbol{\varepsilon}$ is being tested for its adequacy.

Under the assumption that columns $\varepsilon_{\cdot j}$ of the noise matrix $\boldsymbol{\varepsilon}$ are i.i.d. from $\mathcal{N}_p(\mathbf{0}, \Sigma)$ where Σ is a positive definite matrix, the likelihood ratio test (LRT) for $H_0 : \mathbf{B}_1 = \mathbf{B}_{10}$ rejects the null hypothesis for small values of the statistic

$$L_{reg} = \frac{|n\widehat{\Sigma}|}{|n\widehat{\Sigma} + (\widehat{\mathbf{B}}_1 - \mathbf{B}_{10})\mathbf{A}_{11\cdot 2}(\widehat{\mathbf{B}}_1 - \mathbf{B}_{10})^T|} \tag{23}$$

where $\widehat{\mathbf{B}}$ is the submatrix with first k_1 columns of $\widehat{\mathbf{B}}$, the least squares estimator of $\mathbf{B}$, given by $\mathbf{YX}^T\mathbf{A}^{-1}$, $\widehat{\Sigma} = n^{-1}(\mathbf{Y} - \widehat{\mathbf{B}}\mathbf{X})(\mathbf{Y} - \widehat{\mathbf{B}}\mathbf{X})^T$, $\mathbf{A}_{11\cdot 2} = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$, where $\mathbf{A} = \mathbf{XX}^T$ and $\mathbf{A}_{ll'} = \mathbf{X}_l\mathbf{X}_{l'}^T$ for $1 \le l, l' \le 2$. It is well-known that, under the stated distributional assumptions, $\mathbf{G} = n\widehat{\Sigma}$ has a Wishart$(p, n-k; \Sigma)$ distribution. Also, under the null hypothesis, $\mathbf{H} = (\widehat{\mathbf{B}}_1 - \mathbf{B}_{10})\mathbf{A}_{11\cdot 2}(\widehat{\mathbf{B}}_1 - \mathbf{B}_{10})^T$ has a Wishart$(p, k_1; \Sigma)$ distribution, and $\mathbf{G}$ and $\mathbf{H}$ are independent. Thus,

by expressing the statistic L_{reg} as $1/|I_p + \mathbf{H}\mathbf{G}^{-1}|$, we notice that the LRT is equivalent to rejecting H_0 for large values of the statistic

$$\widetilde{L}_{reg} = \log |I_p + \mathbf{H}\mathbf{G}^{-1}|. \tag{24}$$

The latter object can be identified as a linear spectral statistic (LSS) in the *Fisher matrix* $\mathbf{H}\mathbf{G}^{-1}$, under H_0, i.e., a function of $\mathbf{M} = \mathbf{H}\mathbf{G}^{-1}$ of the form $\sum_{j=1}^{p} g(\lambda_j(\mathbf{M}))$, with $\lambda_j(\mathbf{M})$'s denoting the eigenvalues of $\mathbf{M}$ (all real), and $g(x) = \log(1+x)$.

The last observation is the cornerstone for much of the theoretical developments of such inference problems in the RMT framework. In particular, Bai et al. (2013) utilized this representation, and a CLT for LSS of F-matrices, to derive the asymptotic null distribution of the LRT statistic under the assumptions that $p/k_1 \to \kappa_1 \in (0, \infty)$ and $p/(n-k) \to c \in (0, 1)$. Using this, they proposed a correction term in the LRT test to account for large dimensionality of the predictors and the linear constraints (see Sec. 7.6 of Yao, Zheng and Bai (2015) for details). They also conjectured that such results may actually hold for non-Gaussian observations, under the generative model (18) where $\boldsymbol{\varepsilon} = \Sigma^{1/2}\mathbf{Z}$ where $\mathbf{Z}$ is a $p \times n$ matrix with i.i.d. coordinates with zero mean, unit variance and finite fourth moments. Such a result and its generalizations for a class of spectrally regularized tests of linear hypotheses, under appropriate regularity conditions, have recently been derived by Li, Aue and Paul (2020), which we discuss later in Section 5.3.

4.3 Ridge regression

Ridge regression or Tikhonov regularization of the sample covariance matrix (or the Gram matrix) of the predictors has been one of the preferred approaches to dealing with the multicollinearity problem in linear regression involving many predictor variables. Such regularization schemes are known to reduce the variability of the estimator, in comparison with the standard least squares estimator, at the cost of introducing some bias, resulting in the classic *bias-variance trade-off* issue in terms of the choice of the regularization parameter. Until recently, theoretical analyses of such schemes have only been done in the fixed p, or p growing slowly with n, regimes. Much of these analyses have focused on the risk behavior of the ridge estimator of the regression coefficient and cross-validation or AIC-type criteria have been used, with asymptotic justification under fixed p regime, for selection of the ridge regularization parameter.

More recently, such regularization schemes have been analyzed within the RMT framework, i.e., when p is comparable to, or even larger than n, and assuming that predictors are random. To be specific, suppose we are dealing with the univariate regression model

$$Y_i = X_{i\cdot}^T \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{25}$$

with i.i.d., noise ε_i's with zero mean, and p dimensional predictors $X_{i\cdot}$. In vector form, the model can be expressed as

$$\mathbf{Y} = \mathbf{X}^T \boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{Y} = (Y_i)_{i=1}^n$, $\boldsymbol{\varepsilon} = (\varepsilon_i)_{i=1}^n$ and $\mathbf{X}$ is a $p \times n$ matrix, with i-th row equal to $X_{i\cdot}$, a $p \times 1$ vector. Note that, our notation here is unconventional, but we use it to be consistent with the rest of the manuscript in terms of the fundamental structure of the theoretical analyses. Then, the ridge regression estimator of $\boldsymbol{\beta}$ based on model (25), with regularization parameter $\lambda \geq 0$, is given by $\widehat{\boldsymbol{\beta}}_\lambda = (\mathbf{X}\mathbf{X}^T + \lambda I_p)^{-1}\mathbf{X}\mathbf{Y}$, where the inverse always exists as long as $\lambda > 0$. In this formalism, assuming that the columns of $\mathbf{X}$ are independent random realization from some p-dimensional

distribution, it is easy to see that ridge regression actually imposes a shrinkage, determined by the parameter λ, on the eigenvalues of the empirical covariance matrix $n^{-1}\mathbf{X}\mathbf{X}^T$ of the predictor variables (we are assuming for conveniences that the predictors are centered, or have zero mean).

One of the first comprehensive theoretical analyses of the risk characteristics of ridge regression estimators, and the dependence of this behavior on the geometry of the high-dimensional data cloud corresponding to the predictor variables, under model (25) and when $p/n \to c > 0$, was carried out by El Karoui and Kösters (2011). Dobriban and Wager (2018) derived the predictive risk characteristics of ridge regression under the same asymptotic framework, while assuming that the coefficient vector is random. Other works analyzing the risk characteristics of ridge regression in large dimensions include Hsu, Kakade, and Zhang (2014) and Dicker (2013).

Dobriban and Liu (2019) gave a fundamental representation for high-dimensional ridge estimator in the $p/n \to c > 0$ setting. Specifically, assuming that the design matrix $\mathbf{X}$ is of the form $\Sigma^{1/2}\mathbf{U}$, where $\mathbf{U}$ is a $p \times n$ matrix with i.i.d. zero mean, unit variance entries, and $\Sigma^{1/2}$ is the square-root of a positive definite matrix, the noise ε_i's are i.i.d. with zero mean and variance σ^2, independent of $\mathbf{X}$, and that $\limsup \|\boldsymbol{\beta}\|_2 < \infty$, they showed that the ridge estimator $\widehat{\boldsymbol{\beta}}_\lambda$ corresponding to regularization parameter λ, has the following asymptotic representation:

$$\widehat{\boldsymbol{\beta}}_\lambda \asymp A_\lambda(\Sigma)\boldsymbol{\beta} + B_\lambda(\Sigma)\frac{\sigma}{\sqrt{n}}Z \tag{26}$$

where $Z \sim \mathcal{N}_p(\mathbf{0}, I_p)$ stochastically depends on $\boldsymbol{\varepsilon}$, and

$$A_\lambda(\Sigma) = (\gamma_p(\lambda) + \gamma_p'(\lambda))(\gamma_p(\lambda)\Sigma + \lambda I_p)^{-2}\Sigma, \qquad A_\lambda(\Sigma) = \gamma_p(\lambda)(\gamma_p(\lambda)\Sigma + \lambda I_p)^{-1}\Sigma,$$

where $\gamma_p(\lambda) \in (0,1)$ is the unique solution to the equation

$$1 - \gamma_p(\lambda) = \frac{\gamma_p(\lambda)}{n}\mathrm{tr}\left[\Sigma(\gamma_p(\lambda)\Sigma + \lambda I_p)^{-1}\right].$$

The equivalence in (26) holds in the sense that the inner product of the difference between the vectors on both sides with any arbitrary l^2-bounded deterministic weight vector (i.e., an arbitrary linear combination of the difference vector) is asymptotically negligible. Notice that (26) effectively quantifies the asymptotic bias and variability of the ridge estimator. Dobriban and Liu (2019) also studied the bias of the K-fold cross-validation (CV) procedure for choosing the regularization parameter, based on which they proposed a simple bias-correction method for the CV score. Moreover, they also analyzed the accuracy of primal and dual sketching estimators based on ridge regression. The latter method is designed to reduce the computational burden associated with the estimation procedure, especially in the context of massive data sets.

4.4 Robust regression

Robust regression has drawn considerably attention in decades. Consider a design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$, as well as a vector of univariate responses $\mathbf{Y} \in \mathbb{R}^n$. Assume that the univariate linear regression model (25) is satisfied. Under this set up, a robust estimator of $\boldsymbol{\beta}$, say an M-estimator, is defined as

$$\widehat{\boldsymbol{\beta}}_\rho = \arg\min_{\boldsymbol{\beta}\in\mathcal{R}^p} \sum_{i=1}^n \rho\left(Y_i - X_{i\cdot}^T\boldsymbol{\beta}\right). \tag{27}$$

In the classical framework, when p is fixed and $n \to \infty$, Huber (1973) derived that if the design matrix $\mathbf{X}$ is full rank, under additional regularity conditions on $\mathbf{X}$, ρ and "score function" $\psi = \rho'$,

the estimator $\widehat{\boldsymbol{\beta}}_\rho$ is asymptotically Gaussian with mean $\boldsymbol{\beta}^*$ (true value of $\boldsymbol{\beta}$) and

$$\text{Var}(\widehat{\boldsymbol{\beta}}_\rho) = (\mathbf{X}^T\mathbf{X})^{-1}\frac{\mathbb{E}(\psi^2(\varepsilon))}{[\mathbb{E}(\psi'(\varepsilon))]^2} := C(\rho,\varepsilon)(\mathbf{X}^T\mathbf{X}) \tag{28}$$

where ε has the same distribution as the i.i.d. noise ε_i's. Huber and Ronchetti (2009) derived that for p fixed (later generalized in Huber (1973) to a regime of larger p such that $p^2/n \to 0$ and $p^3/n \to 0$ (Portnoy, 1985)), the optimal loss function ρ, when seeking to minimizing (28), is of the form $\rho_{opt} = -\log f_\varepsilon$, where f_ε is the density function of error distribution ε. In other words, this implies that maximum likelihood estimator is optimal.

High dimensional case

When p, n both go to infinity such that $p/n \to c \in (0,1)$, the asymptotic behaviors of estimator $\widehat{\boldsymbol{\beta}}_\rho$ and the risk $\mathbb{E}\|\widehat{\boldsymbol{\beta}}_\rho - \boldsymbol{\beta}^*\|^2$ are qualitatively different from the low-dimensional case. Indeed $\widehat{\boldsymbol{\beta}}_\rho$ is no longer a consistent estimator of $\boldsymbol{\beta}_0$ in the L^2 norm. In the classical framework, for any fixed sequence $a_n \in \mathcal{R}^n$,

$$\frac{a_n^T(\widehat{\boldsymbol{\beta}}_\rho - \boldsymbol{\beta}^*)}{\sqrt{a_n^T\Sigma_n a_n}} \overset{D}{\Longrightarrow} \mathcal{N}(0,1) \tag{29}$$

where $\Sigma_n = \text{Var}(\widehat{\boldsymbol{\beta}}_\rho)$. Even for the least squares estimators, Huber (1973) showed that (29) is impossible for every a_n when $p/n \to c \in (0,1)$.

El Karoui (2018) gave rigorous results on the asymptotic behavior for both unregularized and ridge-regularized high dimensional robust regression estimators under a more general setting, where the predictors are independent realizations from an elliptically symmetric distribution. It should be noted that there exists no universality in this result with respect to the sampling design, and the results were derived for data $X_{i\cdot} = \lambda_i\mathcal{X}_i$ where $\mathcal{X}_i$'s have i.i.d. coordinates with bounded support, zero mean and unit variance, and the scalar quantities λ_i's are i.i.d with bounded support that are independent of $\mathcal{X}_i$'s. The influence of λ_i's is very nonlinear and hence not only the covariance of $X_{i\cdot}$, but indeed the geometry of the data cloud, has an impact on the result. Results given in this paper essentially yield the fact that the regression coefficient vector estimated by a robust regression procedure contains an extra Gaussian noise component which is not informed by canonical conception such as the Fisher information matrix. Another perspective on this result is provided by Donoho and Montanari (2016) by making use of an *Approximate Message Passing (AMP)* algorithm, even though their formulation requires the entries of the design matrix to be i.i.d. Gaussian, which was not assumed in El Karoui et al. (2013).

In the setting that $p/n \to c < 1$ and Gaussian covariates, El Karoui et al. (2013) gave an explicit stochastic representation for the distribution of $\widehat{\boldsymbol{\beta}}_\rho$ and the limit of $\|\widehat{\boldsymbol{\beta}}_\rho - \boldsymbol{\beta}^*\|$, defined as $r_\rho(c)$, can be expressed through two highly non-linear equations in terms the parameter c (equation (30)), design matrix, error distribution as well as the form of the objective function ρ. El Karoui et al. (2013) formulated the problem of extracting information about $\widehat{\boldsymbol{\beta}}_\rho$ from the "normal equation" $\sum_{i=1}^n X_{i\cdot}\Psi(Y_i - X_{i\cdot}^T\widehat{\boldsymbol{\beta}}_\rho) = \mathbf{0}$ where score function $\Psi = \rho'$, as a random matrix problem, and made use of a "leave-one-out" perturbation argument for both the responses and predictors.

The following result (Corollary 1 in El Karoui et al. (2013)) gives the representation of the estimator under a special case when $X_{i\cdot} \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \Sigma)$, with nonsingular Σ.

Theorem 5. *Under regularity conditions on $\{\varepsilon_i\}$ and ρ (a convex function), $\left\|\widehat{\beta}_\rho - \beta_0\right\|$ is asymptotically deterministic, where $\|\cdot\|$ denotes the L^2 norm. Let $r_\rho(c)$ be its deterministic limit as*

$n, p \to \infty$ and let $\widehat{z}_\varepsilon$ be a random variable defined as $\widehat{z}_\varepsilon = \varepsilon + r_\rho(c)Z$ where $Z \sim \mathcal{N}(0,1)$ is independent of ε, and the latter has the same distribution as ε_i's. Then $r_\rho(c)$ is determined by the following coupled equations.

$$\begin{cases} \mathbb{E}\left([prox(c\rho)]'\left(\widehat{z}_\varepsilon\right)\right) & = 1 - c \\ \mathbb{E}\left(\left[\widehat{z}_\varepsilon - prox(c\rho)\left(\widehat{z}_\varepsilon\right)\right]^2\right) & = cr_\rho^2(c) \end{cases} \tag{30}$$

where by definition (Moreau, 1965), for a convex function $f : \mathbb{R} \to \mathbb{R}$, the corresponding proximal function prox(f) is defined as

$$prox(f)(x) = \arg\min_{y \in \mathbf{R}} \left(f(y) + \frac{1}{2}(x-y)^2 \right).$$

Lei, Bickel and El Karoui (2018) established the coordinate-wise asymptotic normality of regression M estimates assuming a fixed design matrix $\mathbf{X}$ under appropriate regularity conditions when $p/n \to c \in (0,1)$. With the assumptions on boundedness of first and second derivative of score function ψ, some technical assumptions on ϵ_i to derive Poincaré inequality for non-Gaussian generalization as well as regularity condition on design matrix X, the results of coordinate-wise asymptotic normality of regression M estimates was achieved though introducing a metric, e.g. Kolmogorov distance and total variation distance inducing the weak convergence topology. The proofs relied "leave-one-out" arguments generalized from the techniques used in El Karoui et al. (2013) as well as second-order Poincaré inequality developed in Chatterjee (2009). In the fixed design settings, the inference only involve the randomness of error, some experimental designs as well as survey sampling with moderately large p, n can be carried out.

For extremely large datasets, distributed computation is increasingly used by practitioners. Dobriban and Sheng (2019) worked on one-step and iteratively weighted parameter averaging in general linear models under data parallelism, by implementing linear regression on several independent machines and taking a weighted average of the estimated parameters. Dobriban and Liu (2018) considered "sketch-and-solve" methods and study the asymptotic behaviors of several quantities measuring the performance of currently existing sketching methods such as random projection methods (Gaussian and i.i.d. projections, uniform orthogonal- Haar - projections, subsample randomized Hadamard transforms) as well as random sampling methods (including uniform, randomized leverage-based, and greedy leverage sampling) for unregularized linear regression. A "large data" asymptotic regime was considered, they firstly reduce the data (X, Y) by sketching, say multiplying a $r \times n$ matrix $\mathbf{S}$ that the sketched data is got by $(\widetilde{X}, \widetilde{Y}) = (\mathbf{S}X, \mathbf{S}Y)$. In this setting, both the dimension and sample size goes to infinity that $p/n \to \kappa \in (0,1)$ and the size r of sketched data is also proportional to n such that $r/n \to c \in (\kappa, 1)$. The performance of variance efficiency (VE) $VE(\hat{\beta}_{\mathcal{S}}, \widehat{\beta}) = \frac{\mathrm{E}\|\widehat{\beta}_{\mathcal{S}} - \hat{\beta}\|^2}{\mathrm{E}\|\hat{\beta} - \hat{\beta}\|^2}$ and other quantities including *relative prediction efficiency (PE)*, *residual efficiency (RE)* and *out-of-sample efficiency (OE)* can be expressed in terms of traces of some appropriated matrices, which limit implicitly involved in certain fixed-point equations from the Marchěnko-Pastur law by using RMT techniques.

Analysis of robust regression under the Huber loss function

One particular example of ρ most commonly used in the context of robust statistics is Huber loss function, given as

$$\rho_\tau(x) = \begin{cases} x^2/2 & \text{if } |x| \le \tau, \\ \tau|x| - \tau^2/2 & \text{if } |x| > \tau. \end{cases} \tag{31}$$

in which ρ takes to grow linearly at infinity and the score function $\psi = sgn(x)\min(|x|, \tau)$. In Huber (1981), Huber proposed $\tau = 1.345\sigma$ (fixed) to retain 95% asymptotic efficiency of the estimator for the normally distributed data, and meanwhile to guarantee the estimator's performance towards arbitrary contamination in a neighborhood of the true model. This default setting has found its use in high dimensional statistics even though the asymptotic efficiency is no longer well defined.

One interesting finding is that in the absence of symmetry of error distribution, the bias induced by the Huber loss is not negligible, and these bias on mostly on intercept. The following proposition (Proposition 3.1 in Wang et al. (2018)) provides an explicit bound on the bias, which complements results in Section 4.9.2 in Maronna et al. (2018).

Proposition 1. *Suppose that the observations Y_i's follow the univariate linear regression model with intercept*

$$Y_i = \beta_0 + X_{i\cdot}^T \boldsymbol{\beta} + \varepsilon_i = Z_i^T \boldsymbol{\theta} + \varepsilon_i \tag{32}$$

holds where $\boldsymbol{\theta}^T = (\beta_0, \boldsymbol{\beta}^T)$ is a $1 \times (p+1)$ vector, $Z_i^T = (1, X_{i\cdot}^T)$, and ε_i are i.i.d., zero mean noise. Assume that $\boldsymbol{\varepsilon} = (\varepsilon_i)_{i=1}^n$ and $\mathbf{X} = (X_{i\cdot})_{i=1}^n$ are independent, and X_{ij}'s are i.i.d. with mean μ and variance σ^2. Suppose further that the function $\alpha \to \mathbb{E}\{\rho_\tau(\varepsilon - \alpha)\}$ has a unique minimizer $\alpha_\tau := \arg\min_{\alpha \in \mathbf{R}} \mathbb{E}\{\rho_\tau(\varepsilon - \alpha)\}$, which satisfies

$$\mathbb{P}(|\varepsilon - \alpha_\tau| \le \tau) > 0. \tag{33}$$

Assume further that $\mathbb{E}(\mathbf{Z}\mathbf{Z}^T)$ is positive definite. Then we have

$$\beta_{0,\tau}^* = \beta_0^* + \alpha_\tau \quad \text{and} \quad \boldsymbol{\beta}_\tau^* = \boldsymbol{\beta}^*. \tag{34}$$

where $(\beta_{0,\tau}^*, \boldsymbol{\beta}_\tau^*)$ is defined as $\arg\min_{(\beta_0, \boldsymbol{\beta})} \sum_{i=1}^n \mathbb{E}\rho_\tau(Y_i - \beta_0 - X_{i\cdot}^T \boldsymbol{\beta})$, i.e., the population analog of the Huber estimator $(\widehat{\beta}_{0,\tau}, \widehat{\boldsymbol{\beta}}_\tau)$. Moreover, α_τ, for $\tau > \sigma$, satisfies the bound

$$|\alpha_\tau| \le \frac{\sigma^2 - \mathbb{E}\{\psi_\tau^2(\epsilon)\}}{1 - \tau^{-2}\sigma^2} \frac{1}{\tau}. \tag{35}$$

4.5 Random effects models

There have been comparatively little theoretical developments in terms of analyzing high-dimensional random effects models from the RMT perspective, although this is changing in recent times. Jiang et al. (2016) considered a class of misspecified linear mixed effects models that is motivated by applications to genome-wide association studies. In particular, they established consistency of the REML estimator of the error variance, and convergence (in probability) of the REML estimator of the variance of the random effects to a limit, that depends on the true variance of the random effects and the proportion of the true nonzero random effects present in the linear mixed model. Dicker and Erdogdu (2017) derived uniform concentration bounds and finite sample multivariate normal approximation results for quadratic forms of large dimensional random vectors, used them in the context of variance components estimation in linear random-effects models. Fan and Johnstone (2019) considered the problems of estimation of high-dimensional variance components under the general multivariate mixed effects model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \sum_{l=1}^k \mathbf{U}_l \boldsymbol{\alpha}_l, \qquad \boldsymbol{\alpha}_l \sim \mathcal{N}(\mathbf{0}, I_{m_l} \otimes \Sigma_l), \quad l = 1, \ldots, k, \tag{36}$$

where $\mathbf{Y} \in \mathbb{R}^{n \times p}$ denotes the p-dimensional responses, $\mathbf{X}$ denotes the design matrix corresponding to the fixed effect $\boldsymbol{\beta}$, $\mathbf{U}_l$, denotes the $n \times m_l$ design matrix associated with the random effects $\boldsymbol{\alpha}_l$, and Σ_l denotes the covariance $p \times p$ covariance matrix of the l-th variance component $\boldsymbol{\alpha}_l$. They studied the eigenvalue distribution of standard MANOVA estimators for the variance component covariance matrices $\{\Sigma_l\}_{l=1}^k$, when the dimensionality of the observations is large and comparable to the number of realizations of each random effect. By making use of operator-valued *free probability theory*, they established that the ESD of such estimators are well approximated by deterministic limit laws whose Stieltjes transforms are characterized by systems of fixed-point equations. In a subsequent work, Fan, Sun and Wang (2019) showed that the eigenvalues and eigenvectors of the estimators of the variance components will exhibit quantifiable asymptotic bias in the same asymptotic framework, when the population covariance matrices associated with the variance components have spiked eigenvalue structures.

5 MANOVA and generalizations

In this section we consider the MANOVA problem in the high-dimensional setting within the RMT framework, i.e., when the dimensionality of the observations are comparable to the sample sizes. We start with the simplest possible version of these problems, often referred to as the two sample test for equality of means, and then discuss some generalizations. We also look into related methods that deal with settings where p could even be an order of magnitude larger than n. We end the section with an overview of some recent developments on using spectral regularization procedures for dealing with the inference problems in MANOVA, and more generally, to perform tests for linear hypotheses in high-dimensional linear models.

5.1 Two sample problems

A classical formulation of the two sample problem is to assume that we have i.i.d. samples from two populations of p-dimensional observations $\mathcal{N}_p(\boldsymbol{\mu}_l, \Sigma_l)$, $l = 1, 2$, and the goal is to test whether these two populations are really the same. In the case where Σ_1 and Σ_2 are both unknown and not necessarily equal, the corresponding hypothesis testing problem for equality of means, viz., $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ against $H_A : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ is traditionally referred to as the multivariate Berhens-Fisher problem. In the special case, when $\Sigma_1 = \Sigma_2 = \Sigma$, say, the testing problem becomes a special case of the MANOVA problem described above. A lot of work have gone into solving this special and more general versions of the two sample problems, including through modern machine learning techniques such as use of random projections and kernel-based formulations.

Here, for clarity and in keeping with the overall theme of this, we shall restrict attention to the setting stated above, while allowing the following generalization in terms of distributional assumptions:

A1 $X_{\cdot j}^{(l)} = \Gamma_l Z_{\cdot j}^{(l)} + \boldsymbol{\mu}_l$, $j = 1, \ldots, n_l$, for all l, where $Z_{ij}^{(l)}$ are i.i.d. with zero mean, unit variance and finite fourth moment, for $p \times p$ matrices Γ_l so that $\Gamma_l \Gamma_l^T = \Sigma_l$.

A2 $n_1/n \to \kappa \in (0, 1)$ as $n = n_1 + n_2 \to \infty$.

A3 $p, n \to \infty$ such that $p/n \to c \in (0, \infty)$.

Assuming $\Sigma_1 = \Sigma_2$, Hotelling (1931) proposed the following generalization of the univariate t test, which rejects H_0 for large values of the T^2-statistic:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\overline{X}^{(1)} - \overline{X}^{(2)})^T \mathbf{S}^{-1}(\overline{X}^{(1)} - \overline{X}^{(2)})$$

where $\mathbf{S}$ is the *pooled sample covariance matrix*

$$\mathbf{S} = \frac{1}{n_1 + n_2 - 2}\left[\sum_{l=1}^{2}\sum_{j=1}^{n_l}(X_{\cdot j}^{(l)} - \overline{X}^{(l)})(X_{\cdot j}^{(l)} - \overline{X}^{(l)})^T\right].$$

Notice that, with $n = n_1 + n_2$, we have $\mathbf{S} = n/(n-2)\mathbf{S}_e$, where $\mathbf{S}_e$ is the within group sample covariance. Under Gaussianity, the distribution of $(n-p-1)T^2/p(n-2)$ is a non-central F distribution with degrees of freedom p and $(n-p-1)$ and noncentrality parameter $\Delta = (n_1 n_2/n)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$.

Hotelling's T^2 test has many pleasing statistical properties. For example, this is equivalent to the likelihood ratio test for $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ (Anderson, 2003). Also, the test is invariant under scaling of the variables through a nonsingular matrix multiplication. However, it was observed that this exact test displays poor power characteristics when p is large, mainly owing to the spreading of sample eigenvalues, and the resultant inaccuracy in using $\mathbf{S}^{-1}$ as a substitute for the population-level scaling matrix Σ^{-1} in the definition of T^2 statistic. Various remedies have been suggested in the literature, starting with Dempster's *non-exact tests* (Dempster, 1958, 1960) that bypassed the estimation of Σ^{-1} by directly focusing on the euclidean distance between the sample means. A comprehensive theoretical analysis of the Hotelling's T^2 test, as well as Dempster's proposals under the RMT framework of $p \propto n$ was first carried out by Bai and Saranadasa (1996). They also proposed a new test that essentially ignored the data-driven scaling inherent in the Hotelling's T^2 procedure. Specifically, the test by Bai and Saranadasa (1996) is based on the statistic $\|\overline{X}^{(1)} - \overline{X}^{(2)}\|^2 - (1/n_1 + 1/n_2)\mathrm{tr}(\mathbf{S})$. By consistently estimating the variance of this statistic, they obtained the final, normalized form of the test statistic as

$$BS_n = \frac{(n_1 n_2/n)\|\overline{X}^{(1)} - \overline{X}^{(2)}\|^2 - \mathrm{tr}(\mathbf{S})}{\sqrt{2(n+1)/n}B_n}$$

where $B_n^2 = n^2(\mathrm{tr}(\mathbf{S}^2) - n^{-1}(\mathrm{tr}(\mathbf{S}))^2)/(n+2)(n-1)$ is an unbiased and consistent estimator of $\mathrm{tr}(\Sigma^2)$. Under $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, and assuming **A1**–**A3**, they proved that $BS_n \overset{D}{\Longrightarrow} \mathcal{N}(0,1)$. Thus an asymptotic level α test rejects the null if $BS_n > z_\alpha$, where z_α is the $(1-\alpha)$ quantile of $\mathcal{N}(0,1)$. Further assuming that $\boldsymbol{\mu}^T\Sigma\boldsymbol{\mu} = o(n^{-1}\mathrm{tr}(\Sigma^2))$, with $\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, and $\lambda_{\max}(\Sigma) = o(\mathrm{tr}\Sigma^2)$, they established that

$$\beta_{BS}(\boldsymbol{\mu}) - \Phi\left(-z_\alpha + \frac{n\kappa(1-\kappa)\|\boldsymbol{\mu}\|^2}{\sqrt{2\mathrm{tr}(\Sigma^2)}}\right) \to 0,$$

where $\beta_{BS}(\boldsymbol{\mu})$ denotes the power function of the asymptotic level-α test based on the statistic BS_n, and Φ is the standard normal c.d.f.

Both Dempster's proposals and the Bai-Saranadasa procedure (henceforth, BS) are somewhat restricted in terms their scope of applicability since the theoretical validation only exists when p is comparable to n. Chen and Qin (2010) greatly improved the domain of applicability by making a modification to BS procedure. Specifically, they observed that in the BS test statistic,

controlling the "diagonal" terms $\sum_{j=1}^{n_l} \|X_{\cdot j}^{(l)}\|^2$, for $l = 1, 2$, in the expansion of $\|\overline{X}^{(1)} - \overline{X}^{(2)}\|^2$ becomes critical and therefore requires restrictive assumptions, even though they do not help in distinguishing between the populations. This observation led them to modify the BS statistic by using the statistic

$$T_n = \frac{1}{n_1(n_1-1)} \sum_{i \neq j}^{n_1} X_{\cdot i}^{(1)^T} X_{\cdot j}^{(1)} + \frac{1}{n_2(n_2-1)} \sum_{i \neq j}^{n_2} X_{\cdot i}^{(2)^T} X_{\cdot j}^{(2)} - \frac{2}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} X_{\cdot i}^{(1)^T} X_{\cdot j}^{(2)}$$

as an unbiased estimator of $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$. Further, under mild additional conditions on the distribution of the observations,

$$\mathrm{Var}(T_n) = \left[\frac{2}{n_1(n_1-1)} \mathrm{tr}(\Sigma_1^2) + \frac{2}{n_2(n_2-1)} \mathrm{tr}(\Sigma_2^2) + \frac{4}{n_1 n_2} \mathrm{tr}(\Sigma_1 \Sigma_2) \right] (1 + o(1)),$$

which led to the test statistic $CQ_n = T_n / \sqrt{\widehat{\mathrm{Var}(T_n)}}$, where $\widehat{\mathrm{Var}(T_n)}$ is a consistent estimator of $\mathrm{Var}(T_n)$ that is obtained by replacing $\mathrm{tr}(\Sigma_1^2)$, $\mathrm{tr}(\Sigma_2^2)$ and $\mathrm{tr}(\Sigma_1\Sigma_2)$ in the leading factor of $\mathrm{Var}(T_n)$ by their ratio-consistent estimates obtained from the data. Notice that, this also means that this test is applicable even if $\Sigma_1 \neq \Sigma_2$. Moreover, Chen and Qin (2010) established asymptotic consistency of their test, and power under local alternatives, even when $p \gg n$ or $n \gg p$, as long as $n, p \to \infty$ and certain regularity conditions are satisfied. It should be noted that, when $\Sigma_1 = \Sigma_2$, and p is comparable to n, the performances of the statistics CQ_n and BS_n for testing $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ are quite similar under both null and alternatives (Li et al., 2020). Some generalizations of CQ procedure were proposed by Hu et al. (2017).

One interesting observation is that, Hotelling's T^2 statistic is based on a form of Mahalanobis distance between the sample means where $\mathbf{S}^{-1}$ is used as a surrogate for the population-level scaling matrix Σ^{-1}, while BS procedure (and its modification due to Chen and Qin (2010)) effectively replaces the scaling matrix by the identity matrix. This suggests that there may be a potential gain in power if one replaces the scaling factor by some modification to $\mathbf{S}^{-1}$, for example a ridge-type matrix $(\mathbf{S} + \lambda I_p)^{-1}$ for some $\lambda > 0$. Such a regularization scheme effectively involves a transformation of the eigenvalues of $\mathbf{S}$. In Section 5.3, we discuss a general class of regularized tests based on transforming the spectrum of the (pooled) sample covariance matrix $\mathbf{S}$ within the RMT framework.

5.2 MANOVA when the number of populations is fixed

We now give overview of the general high-dimensional MANOVA problem. The classical formulation of MANOVA is in terms of testing equality of means of k normal populations $\mathcal{N}_p(\boldsymbol{\mu}_l, \Sigma)$, $l = 1, \ldots, p$ with a common covariance matrix Σ. As we see below, all the classical solutions of this hypothesis testing problem relied on the behavior of the eigenvalues of a matrix of the form $\mathbf{S}_b \mathbf{S}_e^{-1}$ where $\mathbf{S}_b$ is the *between-group* sample covariance matrix and $\mathbf{S}_e$ is the *within-group* sample covariance matrix.

This problem has a long history in statistics. Under the classical fixed p regime, the large-sample asymptotic distribution of the eigenvalues of $\mathbf{M} := \mathbf{S}_b \mathbf{S}_e^{-1}$, which is a *Fisher matrix* or an F-matrix in the terminology introduced earlier, was derived under normality by Hsu (1941). Suigura (1976) gave asymptotic expansions of the distribution of eigenvalues of the F-matrix. As p becomes larger, as in the two sample problem (Bai and Saranadasa, 1996), the accuracy of large sample approximations deteriorates. High-dimensional asymptotic results about the behavior of the eigenvalues pf $\mathbf{F}$, when $p/n \to c \in (0, 1)$ were obtained by Fujikoshi, Himeno and Wakaki (2008)

who assumed that the population eigenvalues are simple, by assuming normality of the populations. Many other tests have been proposed in the framework in which both p and n are large including the possibility that $p > n$. When $p > n$, $\mathbf{S}_e$ is singular, and so the classical invariant tests are not well-defined. Among approaches to address this problem, Srivastava and Fujikoshi (2006) modified the classical tests using generalized MoorePenrose inverse, Srivastava and Kubokawa (2013) introduced a test that is invariant under change of measurement units, and Ullah and Jones (2015) compared performances of different regularized tests for the MANOVA problem. Hu and Bai (2016) provided a comprehensive review of the development of tests of significance under the MANOVA framework.

Bai, Choi and Fujikoshi (2018) dealt with a high-dimensional MANOVA problem in the framework of k (≥ 2) independent populations of p-dimensional observations with a common unknown covariance matrix Σ, assuming that the dimension p is comparable to the sample sizes, but not requiring normality of the populations. They derived the asymptotic joint distributions of the eigenvalues of $\mathbf{S}_b\mathbf{S}_e^{-1}$ under the null case of equality of population means and also under a sequence of local alternatives. The important feature of their result is that the null and and non-null distributions of the eigenvalues and invariant test statistics are asymptotically robust against departure from normality in high-dimensional situations. Their results also imply that the standard tests for the null hypothesis of equality of populations, such as the likelihood ratio (LR) criterion, Lawley–Hotelling (LH) criterion or Bartlett–Nanda–Pillai (BNP) criterion, that are derived under the normality of the populations, can also be applied under the non-normal setting described there. Under a sequence of local alternatives, the asymptotic distributions of the eigenvalues are shown to be that of the eigenvalues of a weighted sum of a Gaussian Orthogonal Ensemble and a fixed positive definite matrix, the latter determined by the local alternatives. Based on these results, Bai, Choi and Fujikoshi (2018) derived the asymptotic powers of the above multivariate tests in closed form. To give a glimpse of these results, and to set up the discussion for the follow-up works involving regularized test procedures, we describe below the set up and state the main results.

The key technical result of Bai, Choi and Fujikoshi (2018) is stated as follows. Let $q = k - 1$. Let $\mathbf{Z} = ((Z_{ij}))$ denote the $p \times n$ matrix consisting of i.i.d. entries with zero mean, unit variance and finite fourth moment. Let $\mathbf{V} = ((v_{ij}))$ be an $n \times q$ matrix satisfying: (i) $\mathbf{V}^T\mathbf{V} = I_{k-1}$; (ii) $\mathbf{V}^T\mathbf{1}_n = \mathbf{0}$; and $\max_{i,j} |v_{ij}| = O(n^{-1/2})$.

Theorem 6. *Suppose $q \geq 1$ is fixed, $p, n \to \infty$ with $p/n \to c \in (0, 1)$. Let $\mathbf{V}$ satisfy (i)–(iii) above. Then*

$$\sqrt{n}\left(\mathbf{V}^T\mathbf{Z}^T\left(\mathbf{Z}\mathbf{Z}^T\right)^{-1}\mathbf{Z}\mathbf{V} - \frac{p}{n}I_q\right) \stackrel{D}{\Longrightarrow} \sqrt{2c(1-c)}\mathbf{W} \tag{37}$$

where $\mathbf{W}$ is a $q \times q$ Gaussian Orthogonal Ensemble (GOE). Also, with $\overline{Z} = n^{-1}\sum_{j=1}^{n} Z_{\cdot j}$, we have

$$\sqrt{n}\left(\mathbf{V}^T\mathbf{Z}^T\left(\mathbf{Z}\mathbf{Z}^T - n\overline{Z}\overline{Z}^T\right)^{-1}\mathbf{Z}\mathbf{V} - \mathbf{V}^T\mathbf{Z}^T\left(\mathbf{Z}\mathbf{Z}^T\right)^{-1}\mathbf{Z}\mathbf{V}\right) \xrightarrow{P} \mathbf{O}_q \tag{38}$$

where $\mathbf{O}_q$ denotes the $q \times q$ matrix of zeros.

Bai, Choi and Fujikoshi (2018) assumed the following generative model for the observations. With l denoting the population index, and $X_{ij}^{(l)}$ denoting a measurement corresponding to the i-th coordinate of the j-th sample observation from the l-th population,

$$X_{\cdot j}^{(l)} = \Sigma^{1/2}Z_{\cdot j}^{(l)} + \boldsymbol{\mu}_l, \qquad j = 1, \ldots, n_l, \quad l = 1, \ldots, k, \tag{39}$$

where $Z_{ij}^{(l)}$ are i.i.d. with zero mean, unit variance and finite fourth moment, and $\Sigma^{1/2}$ is a non-negative definite square-root of Σ. Then

$$\mathbf{S}_b = \frac{1}{n}\sum_{l=1}^{k} n_l(\overline{X}^{(l)} - \overline{X})(\overline{X}^{(l)} - \overline{X})^T \tag{40}$$

where

$$\overline{X}^{(l)} = \frac{1}{n_l}\sum_{j=1}^{n_l} X_{\cdot j} \qquad \text{and} \qquad \overline{X} = \frac{1}{n}\sum_{l=1}^{k}\sum_{j=1}^{n_l} X_{\cdot j}^{(l)} = \frac{1}{n}\sum_{l=1}^{k} n_l \overline{X}^{(l)},$$

and

$$\mathbf{S}_e = \frac{1}{n_l}\sum_{l=1}^{k}\sum_{j=1}^{n_l}\left(X_{\cdot j}^{(l)} - \overline{X}^{(l)}\right)\left(X_{\cdot j}^{(l)} - \overline{X}^{(l)}\right)^T. \tag{41}$$

In order to establish the asymptotic results, they also assumed the following *balance condition* for the sample sizes

$$\alpha_{nl} := \frac{n_l}{n} \to \alpha_l \in (0,1), \qquad \text{as } n \to \infty, \quad \text{for } l = 1, \ldots, k. \tag{42}$$

The main result on the behavior of the eigenvalues of $\mathbf{M} = \mathbf{S}_b\mathbf{S}_e^{-1}$ under the null hypothesis, i.e., $\boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_k$, is as follows.

Theorem 7. *With $q = k-1$, let $\ell_1 > \cdots > \ell_q > 0$ be the ordered nonzero eigenvalues of $\mathbf{S}_b\mathbf{S}_e^{-1}$. Then, with the normalization*

$$\widetilde{\ell}_i = \sqrt{n(1-c)^3/2c}(\ell_i - p/(n-p)), \qquad i = 1, \ldots, q,$$

under the null hypothesis, and the assumptions stated above, the joint asymptotic joint distribution of $(\widetilde{\ell}_1, \ldots, \widetilde{\ell}_q)$ is the same as the joint distribution of the ordered eigenvalues of the GOE $\mathbf{W}$.

We skip the details about the result under the local alternatives, except to refer the reader to Sec. 5 of Bai, Choi and Fujikoshi (2018). We shall revisit the question of power under localized alternatives for a class of regularized test procedures in Section 5.3.

The above description of high-dimensional phenomena associated with MANOVA pertains to the case when the number of populations is fixed, even as the dimensionality of the observations and the sample sizes grow proportionately. There is a different asymptotic regime where the number of populations is also assumed to grow with the dimensionality and sample sizes. This regime, perhaps less commonly encountered in standard statistical practices, can still be dealt with by embedding the problem in the high-dimensional linear regression model analyzed by Bai et al. (2013). Details can be found in Sec. 7.5 of Yao, Zheng and Bai (2015).

5.3 Spectral regularization for tests of linear hypotheses

We now discuss some of the recent approaches based on regularization of the spectrum of the sample covariance matrix of the observations (in case of MANOVA), or noise (in case of multivariate regression) to address the issue of loss of power of standard hypothesis tests for linear hypotheses in these problems, such as the likelihood ratio test, when dimensionality and sample sizes are comparable. Recall that in the simplest setting of two sample test (under assumed equal covariance for the two populations), the test proposed by Bai and Saranadasa (1996) replaced the quadratic

form involving the inverse sample covariance matrix in the Hotelling's T^2 statistic with the identity matrix, while other approaches (Srivastava and Du, 2008; Lopes, Jacob and Wainwright, 2011) have sought to use different estimates of the squared Euclidean distance between (appropriately rescaled) population means. Chen et al. (2011) took a different approach in which they replaced the inverse of $\mathbf{S}$ (pooled sample covariance matrix) in Hotelling's T^2 statistic, which only exists if $p \leq n-1$, where $n = n_1 + n_2$ is the total sample size, by $(\mathbf{S} + \lambda I_p)^{-1}$ for a positive regularization parameter λ. Li et al. (2020) carried out detailed theoretical analysis of this *regularized Hotelling's T^2 (RHT)* statistic when $p, n_1, n_2 \to \infty$ such that $n_1/n \to \kappa \in (0,1)$ and $p/n \to c \in (0,\infty)$. In particular they established asymptotic normality of the normalized RHT statistic:

$$T_{n,p}(\lambda) = \frac{\sqrt{p}(p^{-1}\mathrm{RHT}(\lambda) - \widehat{\Theta}_1(\lambda, c_n)}{\sqrt{2\widehat{\Theta}_2(\lambda, c_n)}}, \text{ with } RHT(\lambda) = \frac{n_1 n_2}{n_1+n_2}(\overline{X}^{(1)} - \overline{X}^{(2)})^T(\mathbf{S}+\lambda I_p)^{-1}(\overline{X}^{(1)} - \overline{X}^{(2)}),$$

where $c_n = p/n$, and $\widehat{\Theta}_1(\lambda, c_n)$ and $\widehat{\Theta}_2(\lambda, c_n)$ are centering and scaling statistics that depend only on the eigenvalues of $\mathbf{S}$ and λ, under subGaussianty of the observations. Moreover, supposing that under the alternative the population mean difference $\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, which can be either fixed or random, behaves in such a way that $\sqrt{n}\boldsymbol{\mu}^T D(\lambda)\boldsymbol{\mu} \to q(\lambda, c)$ as $n, p \to \infty$, (in probability, for stochastic alternatives) then under this sequence of local alternatives, the RHT test has a nontrivial limiting power function. Here, $D(\lambda) = (1 + c\Theta_1(\lambda, c)^{-1}\Sigma + \lambda I_p)^{-1}$ where $\Theta_1(\lambda, c)$ (the limiting version of $\widehat{\Theta}_1(\lambda, c)$) depends on the limiting spectral distribution of $\mathbf{S}$. Based on this, they proposed a data-driven method for selecting the regularization parameter, based on a broad class of stochastic alternatives. They also formulated an adaptive test that combines several values of the regularization parameters together through a maximal statistic.

Li, Aue and Paul (2020) extended the work of Li et al. (2020) even further by considering the problem of testing general linear hypothesis $H_0 : \mathbf{BC} = \mathbf{O}$ where $\mathbf{B}$ is the $p \times k$ regression coefficient matrix, and $\mathbf{C}$ is a $k \times q$ constraints matrix (with $q \leq k$), within the multivariate linear regression model (18) when the dimensionality of the response grows with the sample size, while the number of predictor variables remains fixed. This set up also includes MANOVA with a fixed number of populations as special case, and so the results apply to that setting as well. In particular, they considered the matrix-valued statistic $\mathbf{M}(f) = \widehat{H} f(\widehat{\Sigma})$, where $\widehat{\Sigma} = n^{-1}\mathbf{Y}(I_p - \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X})\mathbf{Y}$ is the least squares estimate of the noise covariance Σ, and $\mathbf{H} = \mathbf{U}(\mathbf{C}^T(\mathbf{X}\mathbf{X})^{-1}\mathbf{C})^{-1}\mathbf{U}^T$, where $\mathbf{U} = n^{-1/2}\mathbf{Y}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{C}$. Finally, for a user-specified function, $f : \mathbb{R}_+ \to \mathbb{R}$, $f(\widehat{\Sigma})$ equals $\mathbf{E}f(\mathbf{L})\mathbf{E}^T$ where $\mathbf{E}\mathbf{L}\mathbf{E}^T$ is the spectral decomposition of $\widehat{\Sigma}$ and for the diagonal matrix $\mathbf{L} = \mathrm{diag}(\ell_i)_{i=1}^p$ consisting of eigenvalues of $\widehat{\Sigma}$, $f(\mathbf{L}) = \mathrm{diag}(f(\ell_i))_{i=1}^p$. Thus, f may be viewed as a shrinkage operator applied to the eigenvalues of $\widehat{\Sigma}$, a special case of which is the ridge-type regularization, which corresponds to $f(x) = 1/(x+\lambda)$. The main result in Li, Aue and Paul (2020) is to show that after appropriate normalization (with $\sqrt{n}$ scaling), under the null hypothesis, the matrix $\mathbf{M}(f)$ converges in distribution to a $k \times k$ Gaussian Orthogonal Ensemble, as $p, n \to \infty$ with $p/n \to c \in (0,\infty)$. Based on this fact, they derive regularized versions of the likelihood ratio test, Bartlett–Nanda–Pillai test and Lawley–Hotelling test, determined by the shrinkage function f, by replacing the F-matrix $\mathbf{H}\widehat{\Sigma}^{-1}$ appearing in those tests by the matrix $\mathbf{M}(f)$, followed by a corresponding normalization. They also determined powers of these tests under appropriate sequences of local probabilistic alternatives. This work also requires less restrictive distributional assumptions on the observations, and in that respect it also generalizes the smooth regularization scheme for Hotelling's T^2 statistic studied by Pan and Zhou (2011).

6 Inference on covariance matrices

Classical multivariate analysis related to the inference on the structure of covariance matrices has broadly focused on addressing the following idealized problems, formulated under Gaussianity of the observations.

- Test for sphericity of a multivariate distribution.
- Testing equality of covariance matrices of two (or more) populations.
- Test of independence between two (or more) multi-dimensional populations.

We shall give a brief overview of the random matrix approach to dealing with these classical problems and their solutions. We shall also briefly review some modern developments in high-dimensional settings that do not strictly utilize the random matrix perspective. In each of these problems, we first look at some classical inferential procedures, and then look at their RMT extensions. The classical solutions that we focus on here are of two kinds: (i) likelihood ratio (LR) tests, (ii) tests based on the largest eigenvalue of a suitable matrix (often referred to as *Roy's largest root test*). The fundamental difference between these two types of tests lies in the fact that, behavior of the LR test statistic, under the null hypothesis, depends on the distribution of the Empirical Spectral Distribution of a Wishart-type or a Fisher-type matrix, while as the name "largest root test" suggests, the latter class of tests rely only on the extreme eigenvalue(s) of corresponding matrices. We now elaborate on these problems.

6.1 Test of sphericity

Suppose have i.i.d. observations $X_1, \ldots, X_n$ from $\mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$, and we are interested in testing $\Sigma = \Sigma_0$, where Σ_0 is a specified positive definite matrix. This problem can be equivalently formulated as testing $H_0 : \Sigma = I_p$, by rescaling the observations by $\Sigma_0^{-1/2}$. The likelihood ratio test (LRT) statistic for testing this null hypothesis rejects for small values of the statistic

$$L_{id} = (e/n)^{pn/2} |\mathbf{A}|^{n/2} \exp(-\mathrm{tr}(\mathbf{A})/2) \tag{43}$$

where $\mathbf{A} = \sum_{j=1}^{n} (X_j - \overline{X})(X_j - \overline{X})^T$ is the unnormalized sample covariance matrix, which has the Wishart$(n-1, p; \Sigma)$ distribution, and $|\mathbf{A}|$ denotes the determinant of $\mathbf{A}$. Notice that the LRT is equivalent to rejecting H_0 for large values of the statistic

$$\widetilde{L}_{id} = \mathrm{tr}(\mathbf{S}) - \log|\mathbf{S}| - p \tag{44}$$

where $\mathbf{S} = (n-1)^{-1}\mathbf{A}$ is the sample covariance matrix of the observations. The latter expression shows that the LRT statistic is equivalent to a linear spectral statistic (LSS): $\mathrm{tr} g(\mathbf{S})$ where $g(x) = -x + \log x + 1$. This observation forms the basis of the analysis of asymptotic behavior and necessary corrections of the LRT test for $H_0 : \Sigma = I_p$ when $p/n \to c \in (0, 1)$, since CLT for LSS involving the sample covariance matrix $\mathbf{S}$ can be used to approximate the sampling distribution of $\widetilde{L}_{id}$.

A more general testing problem, in which the null hypothesis is of the form $H_0 : \Sigma = \sigma^2 I_p$, where $\sigma^2 > 0$ is unknown, is referred to as the *test of sphericity*. The likelihood ratio test (LRT) for this problem (Anderson, 2003) rejects H_0 for small values of

$$L_{sp} = \frac{|\mathbf{A}|^{n/2}}{(\mathrm{tr}(\mathbf{A})/p)^{pn/2}}. \tag{45}$$

Thus the test is equivalent to rejecting the null for large values of the statistic

$$\widetilde{L}_{sp} = \mathrm{tr}(\mathbf{S}) - \log|\mathbf{S}| \tag{46}$$

which is easily recognized as an LSS involving $\mathbf{S}$.

John (1971) proposed and analyzed a different test for sphericity that is invariant under rotation of the coordinates. John's test rejects $H_0 : \Sigma = \sigma^2 I_p$ for large values of the statistic $\mathrm{tr}(\mathbf{S}^2)/(\mathrm{tr}(\mathbf{S}))^2$. This test statistic can be thus be recognized as a smooth function of the bivariate LSS $(\mathrm{tr}(\mathbf{S}), \mathrm{tr}(\mathbf{S}^2))$. This enables usage of CLT for LSS of sample covariances to be used for asymptotic analysis of the criterion under the RMT framework.

6.2 Test of equality of covariance matrices

We present this problem only in the context involving two Gaussian populations, $\mathcal{N}_p(\boldsymbol{\mu}_l, \Sigma_l)$, $l = 1, 2$, for simplicity. This problem is formulated as testing the null hypothesis $H_0 : \Sigma_1 = \Sigma_2$. Assuming we have n_l observations $\{X_j^{(l)}\}_{j=1}^{n_l}$ from the l-th population, the likelihood ratio test (LRT) statistic for this problem rejects H_0 for small values of the statistic

$$L_{eq} = \frac{|\mathbf{S}_1\mathbf{S}_2^{-1}|^{N_1/2}}{|(N_1/N)\mathbf{S}_1\mathbf{S}_2^{-1} + N_2/N|^{N/2}} \tag{47}$$

where $N_l = n_l - 1$, $N = N_1 + N_2$, and $\mathbf{S}_l = N_l^{-1}\sum_{j=1}^{n}(X_j^{(l)} - \overline{X}^{(l)})(X_j^{(l)} - \overline{X}^{(l)})^T$, for $l = 1, 2$. Notice that the LRT statistic is equivalent to the statistic

$$\widetilde{L}_{eq} = -(N_1/N)\log|\mathbf{S}_1\mathbf{S}_2^{-1}| + \log|(N_1/N_2)\mathbf{S}_1\mathbf{S}_2^{-1} + 1| \tag{48}$$

which can be recognized as an LSS involving the F-matrix $\mathbf{M} := \mathbf{S}_1\mathbf{S}_2^{-1}$. Therefore, CLT for LSS of F-matrices (Zheng, 2012) can be applied to study the asymptotic properties of the LRT statistic under the regime that $p, n_1, n_2 \to \infty$ such that $p/n_1 \to c_1 \in (0, 1)$ and $p/n_2 \to c_2 \in (0, 1)$.

6.3 Test of independence of two multivariate populations

Suppose we have $p+q$ dimensional observations X_j, $j = 1, \ldots, n$, from the $\mathcal{N}_{p+q}(\boldsymbol{\mu}, \Sigma)$ distribution that are partitioned as

$$X_j = \begin{bmatrix} X_j^{(1)} \\ X_j^{(2)} \end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

where $X_j^{(1)}$ is $p \times 1$ and $X_j^{(2)}$ is $q \times 1$, with corresponding dimensions for $\boldsymbol{\mu}_l = \mathbb{E}(X_j^{(l)})$, $l = 1, 2$ and $\Sigma_{lm} = \mathrm{Cov}(X_j^{(l)}, X_j^{(m)})$, $1 \le l, m \le 2$. By the characterization of Gaussianity, the two subvectors are independently distributed if and only if the null hypothesis $H_0 : \Sigma_{12} = \mathbf{O}$ holds, where $\mathbf{O}$ is a $p \times q$ matrix of zero entries.

In this case, define $\mathbf{A} = \sum_{j=1}^{n}(X_j - \overline{X})(X_j - \overline{X})^T$ with a corresponding partitioning having components $\mathbf{A}_{lm}$, $1 \le l, m \le 2$. Then, assuming $n - 1 > p + q$ and Σ_{11} and Σ_{22} to be positive definite, so that $\mathbf{A}_{11}$ and $\mathbf{A}_{22}$ are nonsingular with probability 1, the LRT test reduces to rejecting H_0 for small values of the statistic

$$L_{ind} = \frac{|\mathbf{A}|}{|\mathbf{A}_{11}| \cdot |\mathbf{A}_{22}|} = \frac{|\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}|}{|\mathbf{A}_{22}|} = |\mathbf{I}_q - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}|. \tag{49}$$

Notice that the last expression implies that the test statistic can be expressed in terms of eigenvalues of the symmetric $\mathbf{M} = \mathbf{A}_{22}^{-1/2}\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1/2}$. Also note that these eigenvalues are the *squared empirical canonical correlation coefficients* between variables $X^{(1)}$ and $X^{(2)}$. Under $H_0 : \Sigma_{12} = \mathbf{O}$, the squared population canonical correlation coefficients, i.e., the eigenvalues of the matrix $\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}$ are all zero. So, indeed under Gaussianity, the test of independence of the two components of the vectors X can be equivalently seen as a test for the presence of nonzero canonical correlation coefficients.

6.4 Behavior of the likelihood ratio tests in high dimensions

Bai et al. (2009) studied the problem of inference on the covariance matrices of Gaussian populations under one sample (covariance equals to a specified matrix) and two sample framework (equality of two population covariances), and explained the failure of the corresponding likelihood ratio test procedures when the dimension p is large compared to the sample size n. Using central limit theorems for linear spectral statistics of sample covariance matrices and of random F-matrices, they proposed necessary corrections for these LR tests that account for high-dimensional effects. They derived asymptotic distributions of these corrected tests under the null hypotheses.

Wang and Yao (2013) extended the analysis of Bai et al. (2009) for test of sphericity by relaxing the distributional assumptions. In particular, they proposed corrections to the likelihood ratio test and Johns invariant test (John, 1971, 1972) in large dimensions, and derived the asymptotic distribution of these test statistics under the null hypothesis. These results are valid for general population, i.e. not necessarily Gaussian, provided a finite fourth-moment condition on the observations is met. Very precise corrections to the LR statistics for both the test of sphericity and the test of equality of two population covariances, termed a *substitution principle*, that mimics the Gaussian setting, but works even under non-Gaussian observations, was established in Zheng, Bai and Yao (2015). Their proposal substitutes the adjusted sample size $N = n1$ for the actual sample size n in the centering terms of the modified statistics, which results in non-negligible performance improvement for the tests. It may be mentioned that John's test for sphericity has been extended to deal with even in the setting $p/n \to \infty$ by Birke and Dette (2005).

For the test of independence of two sets of variables, the LRT statistic L_{ind} in (49) is equivalent to the statistic

$$\widetilde{L}_{ind} = \log|a_n I_q| - \log|\mathbf{F} + a_n I_q|$$

where $a_n = (n-p-1)/p$ and

$$\mathbf{F} = \frac{1}{p}\mathbf{X}^{(2)}\mathbf{P}_1(\mathbf{X}^{(2)})^T\left(\frac{1}{n-p-1}\mathbf{X}^{(2)}(I_{n-1}-\mathbf{P}_1)(\mathbf{X}^{(2)})^T\right),$$

where $\mathbf{X}^{(l)}$ is the matrix with columns $X_j^{(l)}$, $j = 1, \ldots, n$, and $\mathbf{P}_1 = (\mathbf{X}^{(1)})^T(\mathbf{X}^{(1)}(\mathbf{X}^{(1)})^T)^{-1}\mathbf{X}^{(1)}$ is a projection matrix of rank p. Notice that, under H_0, $\mathbf{X}^{(2)}$ and $\mathbf{P}_1$ are independently distributed. Also, using Gaussianity, $\mathbf{X}^{(2)}\mathbf{P}_1(\mathbf{X}^{(2)})^T$ and $\mathbf{X}^{(2)}(I_{n-1}-\mathbf{P}_1)(\mathbf{X}^{(2)})^T$ are distributed as independent q-dimensional Wishart distributions with degrees of freedom p and $n-p-1$, and the same scaling matrix Σ_{22}. Thus, $\mathbf{F}$ can be recognized as an F-matrix under the null hypothesis. Based on this observation, and using a CLT for F-matrices, Jiang, Bai and Zheng (2013) established the asymptotic null distribution of the LRT statistic L_{ind} in the asymptotic regime where $n, p, q \to \infty$ such that $q/p \to c_s \in (0, \infty)$ and $q/(n-p-1) \to c_r \in (0, 1)$.

Among other notable approaches to the problem of testing equality of variances of two population, Li and Chen (2012) proposed an approach that does not depend on the distribution of the observations and is applicable even when $p/n \to \infty$.

6.5 Tests based on the largest eigenvalue

Roys largest root (Roy, 1957) is a common test statistic in multivariate analysis, statistical signal processing and allied fields. Classical tests for linear hypotheses in MANOVA problems use the eigenvalues of the F-matrix $\mathbf{M} = \mathbf{S}_b\mathbf{S}_e^{-1}$ where $\mathbf{S}_b$ and $\mathbf{S}_e$ denote the between- and within-group sample covariances. Roys largest root test is based on the largest eigenvalue of $\mathbf{M}$, while the likelihood ratio test, Bartely–Nanda–Pillai test and Lawley–Hotelling test depend on the entire set of eigenvalues. Similarly, for dealing with the problem of testing linear hypotheses in large dimensional linear models, one can use Roy's largest root test where the corresponding F-matrix $\mathbf{M}$ is a "ratio" of the matrix representing the contribution from the regression model fit under the null hypothesis, and a matrix estimating the covariance of the noise. A closely version of Roy's largest root test also appears in the context of canonical correlation analysis between two sets of variables with a joint Gaussian distribution, as alluded to in Section 6.3. In large dimensions, different variants of the Tracy-Widom limit laws associated with the *Jacobi Orthogonal Ensemble (JOE)* can be used to approximate the null distribution of Roy's largest root statistic, an overview of which can be found in Johnstone and Nadler (2008). Furthermore, Roy's largest root test, associated with a sample covariance matrix, has been used to test the sphericity hypothesis that $\Sigma = I_p$ for a one sample problem. Again, the null distribution of the statistic can be approximated by the Tracy-Widom law associated with the *Laguerre Orthogonal Ensemble (LOE)* Johnstone (2001).

In these contexts, Roys test is most powerful among the common tests when the alternative is of rank one. For fixed dimension and increasing sample sizes, Kritchman and Nadler (2009) showed asymptotic optimality of Roy's test against rank one alternatives, i.e., when Σ is of the form $I_p + \omega\mathbf{v}\mathbf{v}^T$ for a unit $p \times 1$ vector $\mathbf{v}$ and $\omega > 0$ denoting the signal strength. The latter can be seen as a special case of a "spiked covariance matrix", that has been analyzed extensively in the context of high-dimensional principal component analysis (see, for example, Paul and Aue (2014), Johnstone and Paul (2018) and the references therein). Onatski, Moreira and Hallin (2013) analyzed the power of the likelihood ratio test of sphericity under rank one alternatives, in the $p/n \to c \in (0, 1)$ regime, and established the convergence, under the null hypothesis and contiguous alternatives, to a Gaussian process indexed by the norm of the perturbation.

For MANOVA and general linear hypothesis testing problems, Johnstone and Nadler (2008) derived accurate approximations for the distribution of Roy's test statistic under a rank-one alternative. Their formulation assumes that under the alternative, the noncentrality matrix (the population version of $\mathbf{M}$ in the MANOVA problem) has the form $\omega\mathbf{v}\mathbf{v}^T$ with $\omega > 0$, and $\mathbf{v} \in \mathbb{R}^p$ is an arbitrary and unknown unit norm vector.

7 Discriminant analysis

The classical version of the discriminant analysis problem assumes that we have a p-dimensional observation from one of k possible multivariate Gaussian populations, and the task of discriminant analysis (a.k.a. classification problem) is to decide which population (or class) this observation belongs to. To be more specific, suppose that l-th population has $\mathcal{N}_p(\boldsymbol{\mu}_l, \Sigma_l)$ distribution, for

$l = 1, \ldots, k$. Suppose further that the prior probabilities associated with these classes are $\pi_1, \ldots, \pi_k$ with $\pi_l > 0$ for all l and $\sum_{l=1}^k \pi_l = 1$. Then, the optimal *Bayes classifier* reduces to a *Quadratic Discriminant Analysis* that establishes class boundaries that are piecewise quadratic surfaces as a function of the observation.

For simplicity of exposition of the high-dimensional phenomena associated with the classification problem, we shall restrict attention to the simpler problem, where the different classes have the same variance, i.e., $\Sigma_l = \Sigma$ for all l. In that case, the optimal *Bayes classifier* reduces to a *Linear Discriminant Analysis*, with hyperplane boundaries between classes. In that case, supposing for the time being that the population parameters are known, the Bayes classification rule, or LDA criterion, classifies an observation x into class j if

$$u_{jl}(x) > \log(\pi_l/\pi_j) \text{ for all } 1 \leq l \neq j \leq k, \text{ where } u_{jl}(x) = \left(x - \frac{1}{2}(\boldsymbol{\mu}_j + \boldsymbol{\mu}_l)\right)^T \Sigma^{-1}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_l). \quad (50)$$

Notice that $u_{jl}(x)$ is simply the log of the likelihood ratio between classes j and l.

In practice, we need to estimate $\boldsymbol{\mu}_l$'s and Σ. It is common to use the estimates $\widehat{\boldsymbol{\mu}}_l = \overline{X}^{(l)} = n_l^{-1} \sum_{i=1}^{n_l} X_{\cdot i}^{(l)}$ and $\widehat{\Sigma} = n^{-1} \sum_{l=1}^k \sum_{i=1}^{n_l} (X_{\cdot i}^{(l)} - \overline{X}^{(l)})(X_{\cdot i}^{(l)} - \overline{X}^{(l)})^T$ (MLE for Σ, also equal to the within-group sample covariance $\mathbf{S}_e$), where $n = \sum_{l=1}^k n_l$, and then plug-in these estimates to obtain estimated $u_{jl}(x)$ and then use the LDA rule (50) with $u_{jl}(x)$ replaced by $\widehat{u}_{jl}(x)$. As in similar problems elsewhere, $\widehat{\Sigma}^{-1}$ is a poor estimate of Σ when p is relatively large so that p/n is a non-negligible fraction. This has some implication on the performance of the plug-in rule, often referred to as the "Fisher linear discriminant rule". In a highly influential work, Bickel and Levina (2004) established that the when the number of variables grows faster than the number of observations, a "naive Bayes classifier" which assumes independent covariates greatly outperforms the Fisher linear discriminant rule under broad conditions. This important observation left open the question of possible improvements over the plug-in LDA rule in the RMT regime.

Even though Serdobolskii (1983) is credited with an initial analysis of LDA in the p comparable to n setting, to the best of our knowledge, the first systematic theoretical analysis of the LDA scheme, and a regularized version, proposed by Friedman (1989) and referred to as RDA (Regularized Discriminant Analysis), within the RMT framework, was carried out by El Karoui and Kösters (2011). They established that in the RMT regime, under Gaussianity and for the two-class discrimination problem, the plug-in LDA in general has suboptimal misclassification rate, and suggested modifications to the decision boundary to achieve optimal asymptotic misclassification rate. They also studied geometric sensitivity of the results with respect to the generative model for the data, e.g., when the Gaussianity of the distribution is replace by ellipticity. El Karoui and Kösters (2011) also carried out an analysis of the risk of the RDA procedure proposed by Friedman (1989). RDA simply replaces $\widehat{\Sigma}$ by a ridge regularized estimate $\widetilde{\Sigma}(\lambda) = (1-\lambda)\widehat{\Sigma} + \lambda \mathbf{C}$, where $\mathbf{C}$ is a pre-specified matrix (typically identity, more generally a diagonal matrix, e.g., the diagonal of $\widehat{\Sigma}$), so that depending on the weight $\lambda \in (0,1)$, $\widetilde{\Sigma}(\lambda)$ linearly shrinks towards $\mathbf{C}$. They determined the shrinkage factor λ needed to attain the optimal misclassification rate based on RDA. Using a slightly different approach, Dobriban and Wager (2018) carried out a predictive risk analysis of RDA within the RMT framework.

To understand the characteristics of the high-dimensional phenomena associated with a classification problem, let us consider the 2-class discriminant analysis problem under the standard framework of two normal populations that differ only through their means. We demonstrate the phenomena through an analysis of the D-rule and T-rule proposed by Saranadasa (1993), as presented in Yao, Zheng and Bai (2015). The basic idea behind these formulations is to first note

that, the unnormalized within-group (or pooled) covariance matrix $\mathbf{A}$ for this 2-class problem equals $\mathbf{A} = (n-2)\widehat{\Sigma} = (n-2)\mathbf{S}_e$. If observation x is classified in class l, then the unnormalized within-group covariance will be updated to

$$\mathbf{A}_l = \mathbf{A} + \alpha_l(x - \overline{X}^{(l)})(x - \overline{X}^{(l)})^T, \qquad \text{where } \alpha_l = \frac{n_l}{n_l + 1}, \qquad l = 1, 2.$$

The T-rule classifies x in class j if $\mathrm{tr}(\mathbf{A}_j) = \min_l \mathrm{tr}(\mathbf{A}_l)$, while the D-rule classifies x in class j if $\det(\mathbf{A}_j) = \min_l \det(\mathbf{A}_l)$. These rules simplify to the following:

$$T - \text{rule : classify in class j} \quad \text{if} \quad \alpha_j(x - \overline{X}^{(j)})^T(x - \overline{X}^{(j)}) = \min_l \alpha_l(x - \overline{X}^{(l)})^T(x - \overline{X}^{(l)})$$

$$D - \text{rule : classify in class j} \quad \text{if} \quad \alpha_j(x - \overline{X}^{(j)})^T\mathbf{A}^{-1}(x - \overline{X}^{(j)}) = \min_l \alpha_l(x - \overline{X}^{(l)})^T\mathbf{A}^{-1}(x - \overline{X}^{(l)}).$$

Thus, notice that, up to a multiplicative factor, D-rule classifies according to the Mahalanobis distance of the observation x from the class centroids (estimates of class means $\boldsymbol{\mu}_l$), where the scaling is provided by an estimate of the population covariance matrix Σ. Thus, D-rule has the same invariance property with respect to nonsingular linear transformation enjoyed by the Fisher's linear discriminant rule. In contrast, up to a weighting factor, the T-rule classifies observations according to their Euclidean distance from the class centroids, thereby ignoring the scale determined by Σ. Moreover, D-rule requires non-singularity of Σ and $p > n-2$ so that $\mathbf{A}^{-1}$ exists, while T-rule can be generally applied.

Implications of these differences are apparent in the behavior of the associated misclassification probabilities. To state the corresponding result, we introduce two quantities: with $\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, define

$$\Delta = \sqrt{\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}} \qquad \text{and} \qquad \widetilde{\Delta} = \|\boldsymbol{\mu}\|^2/\sqrt{\boldsymbol{\mu}^T\Sigma\boldsymbol{\mu}}.$$

Theorem 8. *Suppose that in the 2-class classification problem, where the populations are $\mathcal{N}_p(\boldsymbol{\mu}_l, \Sigma)$, $l = 1, 2$, we have $p, n_1, n_2 \to \infty$ such that $p/n \to c \in (0, 1)$ and $n_1/n \to \kappa \in (0, 1)$, where $n = n_1 + n_2$. Let $P_D(2|1)$ and $P_T(2|1)$, denote the probability of classifying an observation belonging to class 1 into class 2, based on the D-rule and T-rule, respectively. Then*

$$\lim_{p,n_1,n_2\to\infty} \left(P_D(2|1) - \Phi(-\Delta\sqrt{1-c}/2)\right) = 0,$$

and

$$\lim_{p,n_1,n_2\to\infty} \left(P_T(2|1) - \Phi(-\widetilde{\Delta}/2)\right) = 0.$$

To compare performances of these two procedures in this RMT asymptotic framework, Yao, Zheng and Bai (2015) introduced a quantity $\rho = \widetilde{\Delta}/\Delta$, which can be seen as bounded by 1, by Cauchy-Schwarz inequality. By making use of Theorem 8 the argued that if $p/n < 1 - \rho^2 - \epsilon$, for some small fixed ϵ, for all p, n, then D-rule is preferable to T-rule (in terms of having smaller misclassification error rate), while the reverse is true if $p/n > 1 - \rho^2 + \epsilon$. This result is intuitive since for comparatively larger sample sizes (smaller p/n ratio), the benefit of scaling the coordinates through $\mathbf{A}^{-1}$ outweighs the cost associated with the extra randomness, while the benefit of scaling dissipates with increasing dimensions compared to sample sizes. This analysis also points to the potential gain in terms of classification accuracy that can be obtained by considering a ridge-type scaling of the coordinates, as in RDA, i.e., where the scaling factor is of the form $\widetilde{\Sigma}(\lambda)^{-1}$ with $\widetilde{\Sigma}(\lambda)$ as described earlier. It should be noted that the phenomena observed here are not limited to the Gaussian case only, and in particular qualitatively similar results about the behavior of the T-rule and D-rule were derived under general populations by Li and Yao (2016).

8 Linear time series

Stationary linear time series, or linear processes, can be viewed as class of linear models with random effects. This is a very popular model to for describing time series data. In classical multivariate time series analysis, autocovariance matrices play an important role in characertizing the behavior of the process. Therefore, significant amount of information can be gained by analyzing their sample counterparts. Because of this, much of the work related to linear processes in high-dimensional settings has focused on understanding the behavior of eigenvalues of the sample covariance and autocovariance matrices. To fix notations, let $\mathbf{X}_n = [X_1, \ldots, X_n]$ be our data matrix such that $(X_t : t \in \mathbb{Z})$ is a p-dimensional linear process, i.e.,

$$X_t = \sum_{\ell=0}^{\infty} \mathbf{A}_\ell Z_{t-\ell}, \qquad t \in \mathbb{Z}, \tag{51}$$

where $(Z_t), t \in \mathbb{Z}$ is p-dimensional i.i.d. real or complex-valued process. Here, we assume that the dimension p can grow with the sample size such that $p/n \to c$, were c is a non-negative constant. The two asymptotic regimes that that been studied are (i) $c > 0$ and (ii) $c = 0$. We first focus on the case that $c > 0$ and review the results concerning the sample covariance matrix of the process, i.e.

$$\mathbf{S} = \frac{1}{n}\mathbf{X}_n\mathbf{X}_n^* = \frac{1}{n}\sum_{t=1}^{n} X_t X_t^*.$$

Recall that spectral analysis of sample covariance matrices in the case of i.i.d. entries has been developed extensively since the pioneering work of Marčenko and Pastur (1967).

Many attempts have been made to relax the independence assumption across rows and columns of the data matrix in high-dimensional settings. Among the earliest efforts, Yin and Krishnaiah (1986) established the existence of the LSD of sample covariance matrices were the columns of $\mathbf{X}_n$ are distributed isotropically. Extensions to products of a non-negative definite matrix $\mathbf{T}$ with a sample covariance matrix when the underlying distribution is isotropic are given in Bai, Yin and Krishnaiah (1986). Later, Bai and Zhou (2008) considered sample covariance matrices of samples with more general dependence structure such that concentration of quadratic forms around their means holds. More precisely, it was shown that, under the high dimensional settings considered above, the Stieltjes transform of the LSD of the sample covariance matrix $\frac{1}{n}\mathbf{X}\mathbf{X}^*$, $s_F(z), z \in \mathbb{C}^+$, satisfies the *Marčenko–Pastur equation*

$$s_F(z) = \int \frac{1}{\tau\,(1 - c - czs_F(z)) - z} d\,H(\tau), \tag{52}$$

where for all k, $\mathbb{E}\bar{X}_{jk}X_{\ell k} = t_{\ell j}$ and for any nonrandom $p \times p$ matrix $\mathbf{B} = (b_{jk})$ with bounded norm, $\mathbb{E}|X_k^*\mathbf{B}X_k - \mathrm{tr}(\mathbf{BT})|^2 = o(n^2)$ such that $\mathbf{T} = (t_{j\ell})$ has uniformly bounded norm and the ESD of $\mathbf{T}$, $F^{\mathbf{T}}$, tends to a non-random probability distribution H. As a consequence, they obtained the LSD of Spearman's rank correlation matrices, sample correlation matrices, and sample covariance matrices from causal AR(1) models. In particular, they showed that, if rows of the matrix $\mathbf{X}_n$ are i.i.d. copies of an AR(1) process

$$Y_t = \phi Y_{t-1} + Z_t, \qquad \text{with } |\phi| < 1,$$

then the LSD of $\mathbf{S}_n$ exists and its Stieltjes transform is given by $m = \underline{m}/c + (1-c)/(cz)$, where $\underline{m}$ is the unique solution in the upper complex plane to the 4-th degree polynomial equation

$$(z\underline{m} + 1)^2(\underline{m}^2 + 2\underline{m}(1 + \phi^2) + (1 - \phi^2)^2) = c^2\underline{m}^2.$$

They also showed that the LSD of $\mathbf{S}_n$ satisfies (52) when the rows of $\mathbf{X}_n$ are independent copies of a linear process with absolutely summable coefficients. In that direction, Jin et al. (2009) obtained explicit forms of the LSD of the uncentered sample covariance matrix generated by a first-order vector autoregressive (VAR(1)) model and a first-order vector moving average (VMA(1)) model, with parameters for these explicit forms being estimated. These results were also established under weaker moments condition, finite second moment, for the innovation terms, by Wang, Jin and Schlemm (2012a). In addition, they found the relationship between the LSDs of population covariance matrices and the power spectral density function of the process, which can be understood as an extensive version of Szegö theorem (Gray, 2009).

These results were further extended by Pfaffel and Schlemm (2012a) to the case that the rows of the data matrix are independent copies of a linear process such that the coefficients of the linear process are decaying fast enough. More precisely, they assumed

$$X_{i,t} = \sum_{j=0}^{\infty} \psi_j Z_{i,t-j}, \qquad i = 1, \ldots, p,\ t = 1, \ldots, n \tag{53}$$

and considered processes such that $\exists\, a,$ and δ s.t. $|\psi_j| \leq a(j+1)^{-1-\delta}, j \geq 0$, in addition to finite forth moment and a Lindeberg-type condition for innovation terms. It should be mentioned that these conditions hold for a large class of time series models including (fractionally integrated) ARMA processes. Under the aforementioned conditions, they derived an equation for the Steiltjes transform of the LSD of the sample covariance matrices of the process and showed that the Steiltjes transform is the unique solution of the equation. Their results show that the E.S.D. of the sample covariance matrices of the process considered in (53) converges to a non-random distribution which only depends on c and the spectral density of the process. They also presented a version of their result that holds if the rows of the matrix $\mathbf{X}_n$ have a piecewise constant spectral density. Under similar settings as Pfaffel and Schlemm (2012a), Yao (2012) presented the equation for the Stieltjes transform of the LSD of the sample covariance matrices of of the linear process (53), where the coefficients of linear process, ψ_j, are assumed to be absolutely summable. Extension of these results were obtained by Banna and Merlevède (2015) where the rows of data matrix $\mathbf{X}_n$ were considered to be independent copies of $g(Z_1, \ldots, Z_n)$ such that $(Z_t : t \in \mathbb{Z})$ is a sequence of i.i.d. real value random variables, $g : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}$ is a measurable function, with $\mathbb{E}[g(Z_1, \ldots, Z_n)] = 0$. The aforementioned framework is very general and it includes many linear and nonlinear processes. More detailed discussion and examples of such processes are discussed in Wu (2005) and Wu (2011). Further generalizations to stationary processes with variables that are not necessarily functions of an i.i.d. sequence were achieved by Merlevède and Peligrad (2016).

Until recently, relatively little was known about the behavior of the sample covariance of large data matrices when the entries are dependent across both rows and columns. Pfaffel and Schlemm (2012b) considered a special data matrix where they partitioned a long observation record of length pn into p segments of length n and filled, consecutively, rows of $\mathbf{X}_n$ with obtained segments. They characterized the LSD of $p^{-1}\mathbf{X}_n\mathbf{X}_n^T$ by an integral equation for its Stieltjes transform, where the observed time series is considered to be a linear process satisfying the same conditions as in Pfaffel and Schlemm (2012a). A broad class of spatio-temporal dependence across both rows and columns can be formulated through the work Hachem, Loubaton and Najim (2006) who studied the LSD of sample covariance matrices corresponding to a $p \times n$ data matrix $\mathbf{X}$, whose entries are given by $X_{ij} = \sigma(j/p, k/n)Z_{ij}$ where Z_{ij}'s are i.i.d. with zero mean and unit variance, and $\sigma : [0,1] \times [0,1] \to (0, \infty)$ is a continuous function which may be referred to as a *variance profile*. Utilizing these results Hachem, Loubaton and Najim (2005) derived the LSD of sample covariance matrices associated

with a rectangular data matrix $\mathbf{X}$ where entries are realizations of a properly rescaled bi-stationary Gaussian random field, i.e., $X_{jk} = \sum_{(r,r')\in\mathbb{Z}^2} g_{r,r'} Z_{r-j,r'-k}$ where $Z_{r,r'}$ are i.i.d. complex Gaussian random variables and $g \in \ell^1(\mathbb{Z}^2)$ is a deterministic, complex-valued, summable sequence. An extension to processes with non-Gaussian innovations Z_{jk} was achieved by Banna, Merlevède and Peligrad (2015) who utilized a generalization of the Lindeberg principle (Chatterjee, 2006) to prove their results.

In the context of multivariate linear processes with non-trivial dimensional and temporal correlations, Liu, Aue and Paul (2015) established extensions of the Marčenko–Pastur-type limit laws for ESDs of symmetrized sample autocovariance matrices for certain subclasses of linear time series. In particular, they considered linear processes defined in (51), where $(Z_t : t \in \mathbb{Z})$ denotes a sequence of i.i.d. p-dimensional random vectors with mean zero unit second moment and finite fourth moment such that the coefficient matrices are simultaneously diagonalizable. In other words, the assumed conditions imply that up to an unknown rotation in the orthonormal/unitary basis $\mathbf{U}$, (the eigenbasis of the covariance matrix of X_t), the rows of the time series $\{\mathbf{U}^* X_t\}$ are uncorrelated linear processes (independent if the process $\{X_t\}$ is Gaussian). They also assumed a type of random effects model for the coefficients, in that there exists a sequence of values $\lambda_1, \ldots, \lambda_p \in \mathbb{R}^m$ (with $m \geq 1$), continuous functions $f_\ell : \mathbb{R}^m \to \mathbb{R}$, such that $\mathbf{U}^* \mathbf{A}_\ell \mathbf{U} = \mathrm{diag}\{f_\ell(\lambda_1), \ldots, f_\ell(\lambda_p)\}$, and the empirical distribution of $\{\lambda_1, \ldots, \lambda_p\}$ converges weakly to a non-random probability distribution F^A. Under this setting, they established the existence of the LSD of symmetrized lag-τ sample autocovariance matrices

$$\mathbf{M}(\tau) = \frac{1}{2n} \sum_{t=1}^{n-\tau} \left(X_t X_{t+\tau}^* + X_{t+\tau} X_t^* \right), \qquad \tau \in \mathbb{N}_0. \tag{54}$$

and showed that the Stieltjes transform of the LSD of $\mathbf{M}(\tau)$ satisfies the following system of integral equations.

$$\mathcal{S}_\tau(z) = \int \left[\frac{1}{2\pi} \int_0^{2\pi} \frac{h(\lambda,\theta)\cos(\tau\theta)}{1 + cK_\tau(z,\theta)} d\theta - z \right]^{-1} dF^A(\lambda),$$
$$K_\tau(z,\theta) = \int \left[\frac{1}{2\pi} \int_0^{2\pi} \frac{h(\lambda,\nu)\cos(\tau\nu)}{1 + cK_\tau(z,\nu)} d\nu - z \right]^{-1} h(\lambda,\theta) dF^A(\lambda).$$

where $h(\lambda,\theta) = |\sum_{\ell=0}^{\infty} e^{\mathbf{i}\ell\theta} f_\ell(\lambda)|^2$ is the spectral density of the univariate linear process with coefficients $\{f_\ell(\lambda)\}_{\ell=0}^{\infty}$ at frequency θ, $c = \lim(p/n)$ as $n, p \to \infty$, and F^A is the LSD of the joint spectral distribution of the coefficient matrices of the linear process. The description of the *joint spectral distribution* F^A entails that the model for $\{\mathbf{U}^* X_t\}$ can be thought of as a random effects model, with the underlying random effects $\{\lambda_k\}_{k=1}^p$ determining the parameters $(\{f_\ell(\lambda_k)\}_{\ell\geq 0})$ of the k-th *coordinate process* of $\{\mathbf{U}^* X_t\}$ for each $k = 1, \ldots, p$. Namdari (2018) utilized the framework of Liu, Aue and Paul (2015), but used a weighted version of the sample periodogram estimators, and an optimization strategy, to estimate the spectral distribution of the coefficients of multivariate ARMA processes with symmetric and simultaneously diagonalizable coefficient matrices.

Existence of LSDs of symmetrized lag-τ sample autocovariance matrices were previously established by Jin et al. (2014), in the special case that the entries of data matrix were assumed to be i.i.d. random variables with bounded $(2+\delta)$ moments, for some $\delta \in (0, 2]$. The moment condition was further relaxed to finite second moment by Bai and Wang (2014). Substantial extensions on the results in Liu, Aue and Paul (2015) were achieved by Bhattacharjee and Bose (2016). They relaxed the Hermitianness and simultaneous diagonalizablity of condition of the coefficient matrices

assumed in Wang, Aue and Paul (2017), and instead assumed a weaker condition that A_j's are norm bounded and "jointly convergent" in terms of their tracial mixed moments, in the sense of Definition 8.13 in Nica and Speicher (2006). Bhattacharjee and Bose (2016) proved that the LSD of any symmetric polynomial in matrices $\{\widehat{\mathbf{\Gamma}}_\tau, \widehat{\mathbf{\Gamma}}_\tau^*\}_{\tau\geq 0}$ exist where $\widehat{\mathbf{\Gamma}}_\tau = \frac{1}{n}\sum_{t=1}^{n-\tau}\left(X_t X_{t+\tau}^*\right)$. Their derivation was combinatorial in nature and relied upon the algebraic method of *free probability theory*. Using this approach, they also provided a general description for the LSDs in terms of a class of freely independent variables.

In the asymptotic regime that $p, n \to \infty$ such that $p/n \to 0$, Wang, Aue and Paul (2017) established the existence of the LSD of the renormalized and symmetrized sample autocovariance matrices of linear processes with simultaneous diagonalizable coefficient matrices. Bhattacharjee and Bose (2019) developed further the spectral analysis of autocovariance matrices of linear processes in this moderately high dimensional regime. Their proof techniques make use of the *free probability theory*. For any finite degree symmetric matrix polynomial $\Pi_{sym}(.)$, they established the almost sure existence of the LSD of

$$\sqrt{np^{-1}}\left(\Pi_{sym}(\widehat{\Gamma}_\tau, \widehat{\Gamma}_\tau^* : \tau \geq 0) - \Pi_{sym}(\Gamma_\tau, \Gamma_\tau^* : \tau \geq 0)\right), \tag{55}$$

and gave a representation of the LSD in terms of polynomials of a class of freely independent variables. Bhattacharjee and Bose (2019) also proposed various statistical testing methods concerning linear processes. In particular, they proposed both graphical methods and also more formal tests of significance, for deciding on the order of MA processes. Their graphical method for determining the order q of an MA process is based on the fact that the LSDs of $\widehat{\Gamma}_\tau + \widehat{\Gamma}_\tau^*$ are all the same at lag τ if $\tau > q$, and are not so for $\tau \leq q$. Hence the histogram of eigenvalues of $\widehat{\Gamma}_\tau + \widehat{\Gamma}_\tau^*$ for first few values of τ can provide information on the model order q.

Acknowledgement

The authors gratefully acknowledge the support and encouragement of the editor Professor Soumendranath Lahiri. Paul's research was partially supported by the NSF grants DMS-1713120, DMS-1811405 and DMS-1915894. Wang's research is partially supported by the NSFC grant 11701509 and the First Class Discipline of Zhejiang - A (Zhejiang Gongshang University - Statistics).

References

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis (3rd edition)*. Wiley-Interscience.

Bai, Z. D., Chen, J. and Yao, J. (2010). On estimation of the population spectral distribution from a high-dimensional sample covariance matrix. *Australian and New Zealand Journal of Statistics*, **52**, 423–437.

Bai, Z. D., Choi, K. P. and Fujikoshi, Y. (2018). Limiting behavior of eigenvalues in high-dimensional MANOVA via RMT. *The Annals of Statistics*, **46**, 2985–3013.

Bai, Z. D., Jiang, D., Yao, J. and Zheng, S. (2009). Corrections to LRT on large-dimensional covariance matrix by RMT. *The Annals of Statistics*, **37**, 3822–3840.

Bai, Z. D., Jiang, D., Yao, J. and Zheng, S. (2013). Testing linear hypotheses in high-dimensional regressions. *Statistics*, **47**, 1207–1223.

Bai, Z. D. and Saranadasa, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statistica Sinica*, **6**, 311–329.

Bai, Z. D. and Silverstein, J. W. (1996). CLT for linear spectral statistics of large-dimensional sample covariance matrices. *The Annals of Probability*, **32**, 535–605.

Bai, Z. D. and Silverstein, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices (2nd edition)*. Springer.

Bai, Z. D. and Wang, C. (2014). A Note on the Limiting Spectral Distribution of a Symmetrized Auto-Cross Covariance Matrix. *https://arxiv.org/pdf/1403.2578.pdf*

Bai, Z. D., Yin, Y. Q., and Krishnaiah, P. R. (1986). On limiting spectral distribution of product of two random matrices when the underlying distribution is isotropic. *Journal of multivariate analysis*, **19**, 189–200.

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

Banna, M., Merlevède, F. (2015). Limiting spectral distribution of large sample covariance matrices associated with a class of stationary processes *Journal of Theoretical Probability*, **28**, 745783.

Banna, M., Merlevède, F., and Peligrad, M. (2015). On the limiting spectral distribution for a large class of symmetric random matrices with correlated entries. *Stochastic Processes and their Applications*, **125**, 2700–2726.

Bhattacharjee, M. and Bose, A. (2016). Large sample behaviour of high dimensional autocovariance matrices. *The Annals of Statistics*, **44**, 598–628.

Bhattacharjee, M. and Bose, A. (2019). Joint Convergence of Sample Autocovariance Matrices When $p/n \to 0$ with Application. *The Annals of Statistics*, **47**, 34703503.

Birke, M. and Dette, H. (2005). A note on testing the covariance matrix for large dimension. *Statistics and Probability Letters*, **74**, 281–289.

Bickel, P. J. and Levina, E. (2004).Some theory of Fishers linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, **10**, 989–1010.

Chen, S. X. and Qin, Y. L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*, **38**, 808–835.

Chen, L., Paul, D., Prentice, R. L. and Wang, P. (2011). A regularized Hotellings T^2 test for pathway analysis in proteomic studies. *Journal of the American Statistical Association*, **106**, 1345–1360.

Chatterjee, S. (2006). A generalization of the Lindeberg principle. *The Annals of Probability*, **34**, 2061–2076.

Chatterjee, S. (2009). Fluctuations of eigenvalues and second order Poincaré inequalities. *Probability Theory and Related Fields*, **143**, 1–40.

Dempster, A. P. (1958). A high dimensional two sample significance test. *The Annals of Mathematical Statistics*, **29**, 995–1010.

Dempster, A. P. (1960). A significance test for the separation of two highly multivariate small samples. *Biometrics*, **16**, 41–50.

Dicker, L. H. (2013). Optimal equivariant prediction for high-dimensional linear models with arbitrary predictor covariance. *Electronic Journal of Statistics*, **7**, 1806–1834.

Dicker, L. H. and Erdogdu, M. A. (2017). Flexible results for quadratic forms with applications to variance components estimation. *The Annals of Statistics*, **45**, 386–414..

Dobriban, E. and Liu, S. F. (2019). Asymptotics for Sketching in least squares regression. *Advances in Neural Information Processing Systems*, 3670–3680.

Dobriban, E. and Liu, S. F. (2019). Ridge regression: structure, cross-Validation, and sketching. *arXiv:1910.02373*.

Dobriban, E. and Sheng, Y. (2019). Distributed linear regression by averaging. *arXiv:1810.00412*.

Dobriban, E. and Wager, S. (2018). High-dimensional asymptotics of prediction: ridge regression and classification. *The Annals of Statistics*, **46**, 247–279.

Donoho, D. and Montanari, A. (2016). High dimensional robust M-estimation: asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, **166**, 935–969.

Donoho, D. L., Gavish, M. and Johnstone, I. M. (2018). Optimal shrinkage of eigenvalues in the spiked covariance model. *The Annals of Statistics*, **46**, 1742–1778.

El Karoui, N. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *The Annals of Statistics*, **36**, 2757–2790.

El Karoui, N. (2013). Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *https://arxiv.org/pdf/1311.2445.pdf*

El Karoui, N. (2018). On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, **170**, 95–175.

El Karoui, N., Bean, D., Bickel, P. J. and Lim, C. (2013). On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, **110**(36), 14557–14562.

El Karoui, N. and Kösters, H. (2011). Geometric sensitivity of random matrix results: Consequences for shrinkage estimators of covariance and related statistical methods. Preprint available at *arXiv:1105.1404*.

Fan, Z. and Johnstone, I. M. (2019). Eigenvalue distributions of variance components estimators in high-dimensional random effects models. *The Annals of Statistics*, **47**, 2855–2886.

Fan, Z., Sun, Y. and Wang, Z. (2019). Principal components in linear mixed models with general bulk. *arXiv:1903.09592*.

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, **84**, 165–175.

Fujikoshi, Y., Himeno, T. and Wakaki, H. (2004). Asymptotic results of a high-dimensional MANOVA test and power comparison when the dimension is large compared to the sample size. *Journal of the Japan Statistical Society*, **34**, 19–26.

Fujikoshi, Y., Himeno, T. and Wakaki, H. (2008). Asymptotic results in canonoical discriminant analysis when the dimension is large compared to the sample. *Journal of Statistical Planning and Inference*, **138**, 3457–3466.

Fujikoshi, Y., Ulyanov, V. V. and Shimazu, R. (2010). *Multivariate Statistics: High-Dimensional and Large-Sample Approximations*. Wiley, New York.

Geronimo, J. S. and Hill, T. P. (2003). Necessary and sucient condition that the limit of Stieltjes transforms is a Stieltjes transform. *Journal of Approximation Theory*, **121**, 54–60.

Gray, R. M. (2009) Toeplitz and Circulant Matrices: A Review. Available at: http://ee.stanford.edu/ gray/toeplitz.html.

Hachem, W., Loubaton, P., and Najim, J. (2006). The empirical distribution of the eigenvalues of a Gram matrix with a given variance profile. *Annales de l'IHP Probabilits et Statistiques*, **42**, 649–670).

Hachem, W., Loubaton, P. and Najim, J. (2005). The empirical eigenvalue distribution of a Gram matrix: from independence to stationarity. *Markov Processes and Related Fields*, **11**, 629–648.

Hastie, T., Tibshirani, R. and Wainwright, M. J.(2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC.

Bühlmann, P. and van de Geer, S.(2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

Hotelling, H. (1931). The generalization of Student's ratio. *The Annals of Mathematical Statistics*, **2**, 360–378.

Hsu, P. L. (1941). On the limiting distribution of roots of a determinantal equation. *Journal of the London Mathematical Society*, **16**, 183–194.

Hsu, D., Kakade, S. M. and Zhang, T. (2014). Random design analysis of ridge regression. *Foundation of Computational Mathematics*, **14**, 569–600.

Hu, J. and Bai, Z. D. (2016). A review of 20 years of naive tests of significance for highdimensional mean vectors and covariance matrices. *Science China Mathematics*, **59**, 2281–2300.

Hu, J., Bai, Z., Wang, C. and Wang, W. (2017). On testing the equality of high dimensional mean vectors with unequal covariance matrices. *Annals of the Institute of Statistical Mathematics*, **69**, 365–387.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, **35**(1), 73–101.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, **1**(5), 799–821.

Huber, P. J. (1981). *Robust Statistics*. Wiley, New York.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistic, 2nd Edition*. Wiley, New York.

Jiang, D. and Bai, Z. D. and Zheng, S. (2013). Testing the independence of sets of large-dimensional variables. *Science China Mathematics*, **56**, 135–147.

Jiang, J., Li, C., Paul, D., Yang, C. and Zhao, H. (2016). On high-dimensional misspecified mixed model analysis in genome-wide association study. *The Annals of Statistics*, **44**, 2127–2160.

John, S. (1971). Some optimal multivariate tests. *Biometrika*, **58**, 123–127.

John, S. (1972). The distribution of a statistic used for testing sphericity of normal distributions. *Biometrika*, **59**, 169–173.

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

Johnstone, I. M. (2008). Multivariate analysis and Jacobi ensembles: Largest eigenvalue, Tracy-Widom limits and rate of convergence. *The Annals of Statistics*, **36**, 2638–2716.

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

Johnstone, I. M. and Paul, D.(2018). PCA in high dimensions: An orientation. *Proceedings of the IEEE*, **106**, 1277–1292.

Jin, B., Wang, C., Bai, Z. D., Nair, K. K., and Harding, M. (2014). Limiting spectral distribution of a symmetrized auto-cross covariance matrix. *The Annals of Applied Probability*, **24**, 1199–1225.

Jin, B., Wang, C., Miao, B. and Lo Huang, M.-N. (2009). Limiting spectral distribution of large-dimensional sample covariance matrices generated by VARMA. *Journal of Multivariate Analysis*, **100**, 2112–2125.

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**, 3930–3941.

Ledoit, O. and Wolf, M.(2015). Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions. *Journal of Multivariate Analysis*, **139**, 360–384.

Ledoit, O. and Wolf, M.(2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, **40**, 1024–1060.

Lei, L. H., Bickel, P. J. and El Karoui, N. (2018). Asymptotics for high dimensional regression M-estimates: fixed design results. *Probability Theory and Related Fields*, **89**, 600–610.

Li, J. and Chen, S. X. (2012). Two sample tests for high-dimensional covariance matrices. *The Annals of Statistics*, **40**, 908–940.

Li, H., Aue, A., Paul, D., Peng, J. and Wang, P. (2020). An adaptable generalization of Hotellings T^2 test in high dimension. *The Annals of Statistics*, (to appear).

Li, H., Aue, A. and Paul, D. (2020). High-dimensional general linear hypothesis tests via non-linear spectral shrinkage. *Bernoulli*, (to appear).

Li, Z. and Yao, J. (2016). On two simple and effective procedures for high dimensional classification of general populations. *Statistical Papers*, **57**, 381–405.

Liu, H., Aue, A., and Paul, D. (2015). On the MarenkoPastur law for linear time series. *The Annals of Statistics*, **43**, 675–712.

Lopes, M. E., Jacob, L. and Wainwright, M. J. (2011). A more powerful two-sample test in high dimensions using random projection. *Advances in Neural Information Processing Systems*, 1206–1214.

Lytova, A. and Pastur, L. (2009). Central limit theorem for linear eigenvalue statistics of the Wigner and the sample covariance random matrices. *Metrika*, **69**, 153–172.

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

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1980). *Multivariate Analysis*. Academic Press.

Maronna, R. A., Marton, R. D., Yohai, V. J., and Salibián-Barrera, M. (2018). *Robust Statistics: Theory and Methods (with R)*. Wiley, New York.

Merlevède, F., and Peligrad, M. (2016). On the empirical spectral distribution for matrices with long memory and independent rows. *Stochastic Processes and their Applications*, **126**, 2734–2760.

Moreau, J.-J. (1965). Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, **93**, 273299.

Muirhead, R. J. (1982). *Aspect of Multivariate Statistical Theory*. Wiley, New York.

Namdari, J. (2018) *Estimation of Spectral Distributions of a Class of High-dimensional Linear Processes*. PhD Thesis. University of California, Davis.

Nica, A. and Speicher, R. (2006). *Lectures on the Combinatorics of Free Probability*. Cambridge University Press.

Onatski, A., Moreira, M. J., and Hallin, M. (2013). Asymptotic power of sphericity tests for high-dimensional data. *The Annals of Statistics*, **41**, 1204–1231..

Pan, G. M. and Zhou, W. (2011). Central limit theorem for Hotellings T^2 statistic under large dimension. *The Annals of Applied Probability*, **21**, 1860–1910.

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

Pfaffel, O. and Schlemm, E. (2012). Eigenvalue distribution of large sample covariance matrices of linear processes. *arXiv:1201.3828*.

Pfaffel, O. and Schlemm, E. (2012). Limiting spectral distribution of a new random matrix model with dependence across rows and columns. *Linear Algebra and its Applications*, **436**, 2966–2979.

Portnoy, S. (1985). Asymptotic Behavior of M Estimators of p Regression Parameters when p^2/n is Large; II. Normal Approximation. *The Annals of Statistics*, **13**(4), 1403–1417.

Rao, C. R. (1948). Tests of significance in multivariate analysis. *Biometrika*, **35**, 58–79.

Rao, C. R. (1952). *Advanced Statistical Methods in Biometric Research*. Hafner Press, New York.

Rao, C. R. (1955). Estimation and tests of significance in factor analysis. *Psychometrika*, **20**, 93–111.

Rao, C. R. (1959). Some problems involving linear hypotheses in multivariate analysis. *Biometrika*, **46**, 49–58.

Rao, C. R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhya, Series A*, **26**, 329–358.

Rao, C. R. (1965). *Linear Statistical Inference and Its Applications*. John Wiley & Sons.

Rao, C. R. (1972). Estimation of variance and covariance components in linear models. *Journal of the American Statistical Association*, **67**, 112–115.

Rao, C. R. (1976). The 1975 Wald Memorial Lectures: Estimation of parameters in a linear model. *The Annals of Statistics*, **4**, 1023–1037.

Roy, S. N. (1957). *Some Aspects of Multivariate Analysis*. Wiley, New York.

Saranadasa, H. (1993). Asymptotic expansion of the misclassification probabilities of D-and A-criteria for discrimination from two high dimensional populations using the theory of large dimensional random matrices. *Journal of Multivariate Analysis*, **46**, 154–174.

Silverstein, J. W. and Bai, Z. D. (1995). On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate Analysis*, **54**, 175–192.

Silverstein, J. W. and Choi, S. I. (1995). Analysis of the limiting spectral distribution of large dimensional random matrices. *Journal of Multivariate Analysis*, **54**, 295–309.

Serdobolskii, V. I. (1983). On minimum error probability in discriminant analysis. *Dokl. Akad. Nauk SSSR*, **27**, 720–725.

Srivastava, M. S. and Du, M.(2008). A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis*, **99**, 386–402.

Srivastava, M. S. and Fujikoshi, Y. (2006). Multivariate analysis of variance with fewer observations than the dimension. *Journal of Multivariate Analysis*, **97**, 1927–1940.

Srivastava, M. S. and Kubokawa, T. (2013). Tests for multivariate analysis of variance in high dimension under non-normality. *Journal of Multivariate Analysis*, **115**, 204–216.

Suigura, N. (1976). Asymptotic expansions of the distributions of the latent roots and latent vectors of the Wishart and multivariate F-matrices. *Journal of Multivariate Analysis*, **6**, 500–525.

Tracy, C. and Widom, H.(1994a) Level spacing distributions and the Airy kernel. *Communications in Mathematical Physics*, **159**, 151–174,.

Tracy, C. and Widom, H.(1994b) Fredholm determinants, differential equations and matrix models. *Communications in Mathematical Physics*, **163**, 33–72.

Ullah, I. and Jones, B. (2015). Regularised MANOVA for high-dimensional data. *Australian and New Zealand Journal of Statistics*, **57**, 377–389.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press.

Wang, L.,Aue, A. and Paul, D. (2017). Spectral analysis of sample autocovariance matrices of a class of linear time series in moderately high dimensions. *Bernoulli*, **23**, 2181–2209.

Wang, C., Jin, B., and Miao, B. (2011). On limiting spectral distribution of large sample covariance matrices by VARMA(p, q). *Journal of Time Series Analysis*, **32**, 539–546.

Wang, Q. and Yao, J. (2013). On the sphericity test with large-dimensional observations. *Electronic Journal of Statistics*, **7**, 2164–2192.

Wang, L., Zheng, C., Zhou, W. and Zhou, W.-X. (2018). A new principle for tuning-free Huber regression. *Preprint*.

Widom, H.(1999) On the relation between orthogonal, symplectic and unitary ensembles. *Journal of Statistical Physics*, **94**, 347–363.

Wigner, E. P. (1958). On the distribution of the roots of certain symmetric matrices. *Annals of Mathematics*, **67**, 325–328.

Wishart, J. (1928). The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, **20A**, 32–52.

Wu, W. B. (2005). Nonlinear system theory: another look at dependence. *Proceedings of the National Academy of Sciences*, **102**, 1415014154.

Wu, W. B. (2011) Asymptotic theory for stationary processes. *Statistics and its Interface*, **4**, 207226.

Yao, J.-F. (2012). A note on a Marčenko-Pastur type theorem for time series. *Statistics & Probability Letters* **82**, 22–28.

Yao, J., Zheng, S. and Bai, Z. (2015). *Large Sample Covariance Matrices and High-dimensional Data Analysis*. Cambridge University Press.

Yin, Y. Q. and Krishnaiah, P. R. (1986). Limit theorem for the eigenvalues of the sample covariance matrix when the underlying distribution is isotropic. *Theory of Probability and Its Applications*, **30**, 861–867.

Zheng, S. (2012). Central limit theorems for linear spectral statistics of large dimensional F-matrices. *Ann. Inst. H. Poincar Probab. Statist.*, **48**, 444–476.

Zheng, S., Bai, Z. and Yao, J. (2015). Substitution principle for CLT of linear spectral statistics of high-dimensional sample covariance matrices with applications to hypothesis testing. *The Annals of Statistics*, **43**, 546–591.

Zheng, S., Bai, Z. and Yao, J. (2017). CLT for eigenvalue statistics of large-dimensional general Fisher matrices with applications. *Bernoulli*, **23**, 1130–1178.