Central limit theorems for linear spectral statistics of large regularized covariance matrices

Delft University of Technology

Wouter Versteegh (4595947)

July 2020

Abstract

It has been long observed that the sample covariance matrix S_n is a poor estimator of the true covariance matrix Σ_n when the dimension p of the data is of comparable size to the sample size n. In this thesis we shall consider sample covariance matrices S_n in the case when the dimension of the data increases with the sample size to a fixed ratio c as $(p, n) \to \infty$

We will derive a new statistic $p\hat{\alpha}^*$ based on the general linear shrinkage estimator by Bodnar et al. (2014)[1], where $\hat{\alpha}^*$ is the optimal shrinkage quantity. We will show that the new statistic is normally distributed under the null hypothesis $\Sigma_n = I$, where we assume the existence of the fourth moment of our data.

Furthermore, we will do simulation study that compares our new statistic to tests from finite dimensional statistics that have been altered to work in high dimensional statistics by Wang and Yao [3]. We will look at three different hypothesis, the equicorrelation case, the auto-regressive case and a fixed ratio case.

After that, we will look at the non-linear shrinkage estimator based on the work by Ledoit and Peche (2011) [11], and show that, under the null hypothesis, constructing a test is not directly possible like it is in the linear case.

Preface

This thesis has been written as my final project to fulfill my requirements for the Bachelor Applied Mathematics and has been supervised by Dr.Parolya on behalf of the Statistics Department, faculty EEMCS of the Delft University of Technology.

The thing I absolutely liked most about the project is that is truly felt like a *bachelors* thesis, in the sense that a majority of my bachelor courses were somehow relevant for my thesis. When I saw in the original abstract that you would need a bit of Complex Analysis combined with Probability and Statistics, I was intrigued immediately. I had never expected to use Complex Analysis in a project from Statistics.

During the project I discovered a rich and beautiful world of mathematics behind something I was unaware of. Suddenly I had to recall my knowledge from matrices, combine it with complex analysis to derive a probability distribution for new statistics, it was all in one package which I absolutely loved.

That said I do have to admit that sometimes I struggled a bit with the project. These times with COVID-19 are times as we have never seen before, and my initial approach for the project, was suddenly not possible anymore. The university was always my place where I would study and work the best. That affected me a decent bit and adapting to it took a bit more time than I thought.

Fortunately, the mathematics behind the project sparked my interest time and time again, and the project finally allowed me to combine all the knowledge I had collected over the years.

I want to sincerely thank Nestor Parolya for supervising me throughout the project, and showing me the rich and beautiful mathematics that this project had to offer. Even though we never had the opportunity to meet due to the virus, the weekly Zoom calls we had were always very useful and would always give me new insight in what I should do next, if I got stuck on something. Nestor, thank you.

Lastly, I want to thank Dr. Kurowicka for taking seat in my assessment committee.

Wouter Versteegh, July 2020

Contents

Chapter 1

Mathematical Background

Since we will concern ourselves with random matrices, we will need some results from probability theory, statistics and matrix algebra. We will also use a bit of complex analysis to compute certain integrals. In this section we will lay some mathematical foundations by giving definitions, expressions and theorems that we will use throughout this paper. At the end of each section, there will be a small concluding list of the most important things to take away from that section.

1.1 Matrix algebra

We begin the section by stating some definitions about matrices. We then look into some theorems that deal with diagonalizability, and conclude with some theorems about the trace of a matrix.

Definitions

Definition 1.1.1. A $p \times p$ matrix A is *invertible* if there exists a matrix C such that

$$AC = CA = I$$

Here I is the identity matrix. We note $C = A^{-1}$

Definition 1.1.2. A matrix D is a *diagonal* matrix if for each element $[d_{i,j}] = 0$ for each $i \neq j$. That is, on the diagonal of D, there may be non-zero elements, but outside the (main) diagonal each element is 0.

Definition 1.1.3. An $p \times p$ matrix A is *symmetric* if $A = A^T$, where A^T is the transpose of A, where each element of $[A^T_{i,j}] = [A_{j,i}]$

Definition 1.1.4. A $p \times p$ matrix A is *diagonalizable* if there exist an invertible matrix C such that

$$C^{-1}AC = D$$

Here D is a diagonal matrix.

Definition 1.1.5. A $p \times p$ real symmetric matrix A is *positive semi-definite* if for all $x \in \mathbb{R}^p$:

$$x^T Ax \geq 0$$

If it is strictly larger, then it is *positive definite.*

Definition 1.1.6. An *eigenvalue* of a matrix A is a value λ such that $Av = \lambda v$, where v is the corresponding eigenvector.

Definition 1.1.7. The *trace* of a square $p \times p$ matrix A is $tr(A) = \sum_{i=1}^{p} A_{i,i}$, with $A_{i,i}$ the element on the i-th row and i-th column. The trace is then just simply the sum of the diagonal of the matrix.

Definition 1.1.8. The *Frobenius norm* of a matrix $m \times p$ matrix A is

$$||A||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{p} |a_{i,j}|^2} = \sqrt{tr(AA^T)}$$

Here A^T denotes the transpose of A. Note that for symmetric matrices $A = A^T$, the Frobenius norm is $\sqrt{tr(A^2)}$, and thus the squared Frobenius norm $||A||_F^2$ is just $tr(A^2)$.

Definition 1.1.9. The *trace norm* of a matrix A is $tr(\sqrt{AA^T})$. Note that for symmetric matrices $A = A^T$, the trace norm is just the trace of A, since by definition $\sqrt{A^2} = A$

Matrix theorems

A fundamental result about symmetric matrices is the following Theorem 6.8 from Fraleigh, Beauregard (1995) [5]:

Theorem 1.1.1. *Every real $p \times p$ symmetric matrix A is diagonalizable, i.e. there exist diagonal matrix D and invertible C, both $p \times p$ matrices, such that*

$$A = CDC^{-1}$$

Consequently, we have that $AC = CD$

Another very useful theorem, Theorem 5.2 from [5] states that the matrices C and D from above have to do with eigenvalues and eigenvectors:

Theorem 1.1.2. *Let A be an $p \times p$ matrix and let $\lambda_1, \dots, \lambda_p$ be possibly complex scalars and $v_1, \dots, v_n$ be non-zero vectors in n-space, either $\mathbb{R}^n$ or $\mathbb{C}^n$. Let C be the $p \times p$ matrix with the vectors v_i as columns $C = [v_1 v_2 \dots v_n]$ and D be the $p \times p$ diagonal matrix with the values λ_i on the diagonal, 0 elsewhere.*

Then $AC = CD$ if and only if $\lambda_1, \dots, \lambda_p$ are the eigenvalues of A and each v_i is an eigenvector of A corresponding to λ_j, with $j = 1, 2, \dots, n$.

In other words, if we have a real symmetric matrix A, by Theorem 1.1.1 there exists matrices C invertible and D diagonal such that $A = CDC^{-1}$. By Theorem 1.1.2, these matrices consist of eigenvectors (in the case of C) and of eigenvalues on the diagonal in the case of D.

A nice property that also follow are that the eigenvalues are all real, Theorem 5.5 from Fraleigh, Beauregard (1995)[5].

Theorem 1.1.3. *Every real symmetric matrix A is real diagonalizable. That is, if A is symmetric and has has real entries then all it's eigenvalues are real numbers.*

In particular, if our matrix A is real symmetric and positive definite, then it has only positive eigenvalues.

Theorem 1.1.4. *Let A be a real symmetric positive definite matrix. Then the eigenvalues of A are all positive.*

Proof. Let λ be an eigenvalue of A. If $\lambda = 0$, then there exists some eigenvector v such that $Av = 0$. But that would mean that $v^T Av = 0$. This contradicts with A being positive definite.

Now assume $\lambda < 0$, then there exists an eigenvector x such that $Ax = \lambda x$. But if we then multiply with x^T on the left side we get that $x^T Ax = \lambda |x|^2$. This smaller than 0, since λ is negative and the norm $|x|^2 > 0$. This also contradicts with the assumption that A was positive definite. This implies that all eigenvalues of A are positive. □

The main reason why we are interested if a matrix is diagonalizable is that it becomes really easy to take functions of matrices, if we have an analytic function $f(x) = \sum_{n=0}^{\infty} a_n x^n$. For that we first state a lemma:

Lemma 1.1.5. *Let $A = C^{-1}DC$ be an diagonalizable matrix. Then for all $k \in \mathbb{N}, A^k = C^{-1}D^kC$.*

Proof.

$$\begin{aligned}A^k &= (C^{-1}DC)^k \\ &= (C^{-1}DC)(C^{-1}DC)\dots(C^{-1}DC) \text{ (k times)} \\ &= C^{-1}D(CC^{-1})D(CC^{-1})\dots(CC^{-1})DC \\ &= C^{-1}D\dots DC \\ &= C^{-1}D^kC\end{aligned}$$

□

Using this lemma, you can prove that for analytic f, this function applied to a diagonalizable matrix A is a product of the same eigenvectors with transformed eigenvalues.

Theorem 1.1.6. *Let $A = C^{-1}DC$ be a diagonalizable matrix, and f be an analytic function $f(x) = \sum_{n=0}^{\infty} a_n x^n$. Then*

$$f(A) = f(C^{-1}DC) = C^{-1}f(D)C$$

Proof. We have that

$$\begin{aligned}f(A) &= \sum_{n=0}^{\infty} a_n(A)^n \\ &= \sum_{n=0}^{\infty} a_n(C^{-1}DC)^n \\ &= \sum_{n=0}^{\infty} a_nC^{-1}D^nC \text{ (by the lemma)} \\ &= C^{-1}(\sum_{n=0}^{\infty} a_n(D)^n)C \\ &= C^{-1}f(D)C\end{aligned}$$

At the end we made use of the left and right distributivity of matrices. □

This Theorem has some very important implications. First and foremost, it implies that if A is diagonalizable, then $f(A)$ is also diagonalizable with the same matrix C of eigenvectors, and diagonal matrix $f(D)$. Since D is a diagonal matrix, every function we take of it applies to the diagonal element only:

$$D = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & \vdots \\ \vdots & \dots & \ddots & 0 \\ 0 & \dots & \dots & \lambda_p \end{pmatrix}, f(D) = \begin{pmatrix} f(\lambda_1) & 0 & \dots & 0 \\ 0 & f(\lambda_2) & \dots & \vdots \\ \vdots & \dots & \ddots & 0 \\ 0 & \dots & \dots & f(\lambda_p) \end{pmatrix}$$

So if A is diagonalizable with eigenvalues $\lambda_1 \dots \lambda_p$, then $f(A)$ has eigenvalues $f(\lambda_1), \dots f(\lambda_p)$, with the same eigenvectors as A. In particular, the matrix A^2 has eigenvalues $\lambda_1^2, \dots, \lambda_p^2$.

Trace theorems

Furthermore, we will need some theorems about the trace of a matrix A. Recall that we defined the trace as the sum of the diagonal. A handy lemma from page 77 from [6] about the trace of a matrix is that the trace of 2 square matrices is the same, no matter in what order you multiply them

Lemma 1.1.7. *Let A and B both be $p \times p$ square matrices. Then*

$$tr(AB) = tr(BA)$$

Proof. We have

$$\begin{aligned} tr(AB) &= \sum_{i=1}^{p}(AB)_{i,i} = \sum_{i=1}^{p}\sum_{j=1}^{p} A_{i,j}B_{j,i} \\ &= \sum_{i=1}^{p}\sum_{j=1}^{p} B_{j,i}A_{i,j} = \sum_{j=1}^{p}(BA)_{j,j} = tr(BA) \end{aligned}$$

□

We use this lemma to prove that the trace of an analytic function f applied to a diagonalizable matrix A is equal to the sum of it's eigenvalues, where you apply f to the eigenvalues.

Theorem 1.1.8. *Take an $p \times p$ diagonalizable matrix $A = CDC^{-1}$with eigenvalues $\lambda_1, \dots, \lambda_p$, and let $f(x) = \sum_{k=0}^{\infty} a_k x^k$ be analytic function.*
Then

$$tr(f(A)) = \sum_{i=1}^{p} f(\lambda_i)$$

Proof. We have that

$$\begin{aligned} tr(f(A)) &= tr(Cf(D)C^{-1}) = tr((Cf(D))C^{-1}) \\ &= tr(C^{-1}(Cf(D))) = tr((C^{-1}C)f(D)) \\ &= tr(I \times f(D)) = tr(f(D)) = \sum_{i=1}^{p} f(\lambda_i) \end{aligned}$$

□

Here we made use of the associative property of matrix multiplication, the lemma 1.1.7 about the cyclic property that allowed us to switch matrices $Cf(D)$ and C^{-1}, and made use of the fact that $C^{-1}C = I$.

So far we have that: if A is diagonalizable, it can be written as a product of C and D, and we can take analytic functions of it. We conclude the section with a final proposition.

Proposition 1.1.1. *Let A be real diagonalizable $p \times p$ matrix, with eigenvalues $\lambda_1 \dots \lambda_p$. Then*

$$tr(A) = \sum_{i=1}^{p} \lambda_i$$

Proof. This is just a special case of Theorem 1.1.8, with $f(x) = x$ □

Above proposition is actually in general true for all square matrices A, no matter if they are diagonalizable or not. However, since we are only interested in the sample covariance matrix which is symmetric (and thus diagonalizable), we state the results specifically for diagonalizable matrices.

Summary

- If a matrix A is symmetric, it can be written in a form $A = C^{-1}DC$ where C has the eigenvectors as columns, and D has the eigenvalues on the diagonal.
- If f is an analytic function, and A and $p \times p$ matrix with eigenvalues $\lambda_1, \dots, \lambda_p$, then $tr(f(A))$ is equal to $\sum_{i=1}^{p} f(\lambda_i)$

1.2 Probability

Definitions

Before we can do data analysis, we need to establish some ground rules for our data. We will assume that our data lives on some probability triple $(\Omega, F, \mathbb{P})$, where Ω is the space of possible outcomes, F a σ-algebra on Ω, and $\mathbb{P} : \Omega \to [0, 1]$ a proper probability measure.

We also need some notions of convergence of random variables:

Definition 1.2.1. A sequence of random variables $\{X_n\}$ *converges almost surely* towards X if

$$\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1$$

Definition 1.2.2. A sequence of random variables $\{X_n\}$ *converges in probability* to a random variable X if for all $\epsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0$$

Definition 1.2.3. A sequence of random variables $\{X_n\}$ with corresponding cumulative distribution functions $F_n(x)$ *converges in distribution* to a random variable X with cumulative distribution function $F(x)$ if

$$\lim_{n\to\infty} F_n(x) \xrightarrow{D} F(x)$$

Remark: This only holds at the points x where $F(x)$ is continuous.

Definition 1.2.4. Given a sequence of random variables $X_1, \ldots, X_p$, the *covariance matrix* C is the $p \times p$ matrix with elements

$$c_{i,j} = cov(X_i, X_j) = \mathbb{E}(X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j))$$

A central object in this paper will be the sample covariance matrix. The sample covariance matrix is the usual estimator for the true underlying covariance matrix of the data set.

Definition 1.2.5. Given a $p \times n$ data matrix $Y = (\mathbf{y_1}, \ldots, \mathbf{y_n})$, with each $\mathbf{y_i}$ a data vector of dimension p, the *sample covariance matrix* S_n is the matrix

$$S_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{y_i} \mathbf{y_i}^T = \frac{1}{n} Y Y^T$$

Each particular element $s_{i,j}$, which estimates the covariance between the $i-th$ and $j-th$ variable, can be individually calculated by $s_{i,j} = \frac{1}{n} \sum_{k=1}^{n} y_{k,i} y_{k,j}$

This formula assumes that the mean of the data is known. If the mean is unknown, one needs to replace the $\frac{1}{n}$ by $\frac{1}{n-1}$. This is called the substitution principle. However, in our paper we will assume that we know the mean and variance of the data, so we can use the simpler formula with $\frac{1}{n}$.

Definition 1.2.6. Let S be a $p \times p$ matrix with eigenvalues $\{\lambda_1, \ldots \lambda_p\}$. The *empirical distribution function* $F_p(x)$ of the matrix S is

$$F_p(x) = \frac{1}{p} \sum_{i=1}^{p} 1_{(\lambda_i < x)}$$

Here 1_A denotes the indicator function.

The following definition from Yao et al (2015)[2] will be important later, for when we define linear spectral statistics.

Definition 1.2.7. Let S be a $p \times p$ matrix with eigenvalues $\{\lambda_1, \ldots \lambda_p\}$. The *empirical spectral distribution function* F^S of the matrix S is

$$F^S = \frac{1}{p} \sum_{i=1}^{p} \delta_{\lambda_i}$$

Here δ_a denotes the Dirac mass placed at a point a.

Distributions of interest

Three distributions that will be used throughout this paper are the well known normal or Gaussian distribution, the Gamma distribution and the Marcenko-Pastur distribution (or MP).
The first is important because it appears in the classical one-dimensional central limit theorem, but also in the Central Limit Theorem that we will use for our research. The second one won't be used as much, but will be serve as an alternative distribution for us to test with. We will only use it to compare empirical sizes behaviour. The third one is important because it tells us how eigenvalues of the sample covariance matrix are distributed in the limit, see section 1.2.

Definition 1.2.8. A random variable X is normally distributed with parameters $\mathbb{E}[X] = \mu$ and $Var[X] = \sigma^2$ if it has density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma}}$$

If we have $\mu = 0$ and $\sigma^2 = 1$, we call X standard normal.

Definition 1.2.9. A random variable X is Gamma $\Gamma(a, b)$ distributed with shape a and rate b if it has density

$$f(x; a, b) = \frac{b^a x^{a-1} e^{-bx}}{\Gamma(a)}, x \geq 0$$

Here the $\Gamma(a)$ in the denominator is the Gamma-function, and satisfies for integers $n : \Gamma(n) = (n-1)!$. The distribution $\Gamma(a, b)$ has expectation $\frac{a}{b}$ and variance $\frac{a}{b^2}$.

Marcenko Pastur as the limit distribution

This section is dedicated to elaborate what the Marcenko-Pastur distribution is, and why we are interested in this particular distribution.

Definition 1.2.10. We say that a random variable X is Marcenko-Pastur distributed with index $c > 0$ if it has density function

$$p_c(x) = \begin{cases} \frac{1}{2\pi c x}\sqrt{(b-x)(x-a)} & a \leq x \leq b \\ 0 & \text{elsewhere} \end{cases}$$

Here $a = (1-\sqrt{c})^2$ and $b = (1+\sqrt{c})^2$. If it happens that $c > 1$, we place mass $1 - \frac{1}{c}$ on 0. Below in figure 1.1 you can view for certain values of c how the density function looks like.

From Proposition 2.13 from the book by Yao et al. (2015) [2], the moments of the Marcenko-Pastur law are:

Proposition 1.2.1. *Let X be Marcenko Pastur distributed. Then moments $\mu_k = \mathbb{E}[X^k]$ satisfy*

$$\mu_k := \int x^k p_c(x)dx = \sum_{r=0}^{k-1} \frac{1}{r+1}\binom{k}{r}\binom{k-1}{r}c^r$$

In particular, it's expectation is 1, and it's variance is 1+c.

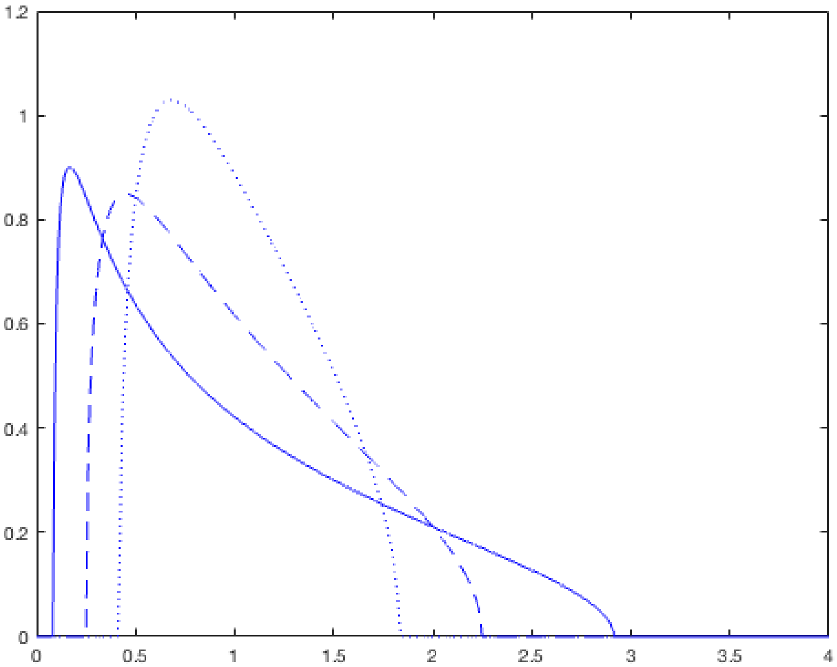


Figure 1.1: Marcenko-Pastur density plotted with different values of c.
c = 1/8 (dotted), c = 1/4 (dashed), c = 1/2 (solid)

The reason why we are interested in this distribution in particular is that the distribution function of the Marcenko-Pastur distribution is the limit of the empirical spectral distribution function of the eigenvalues of a sample covariance matrix S_n. Since the sample covariance matrix itself is a random matrix, it's eigenvalues are random variables. It turns out however that in high-dimensional statistics these eigenvalues obey the Marcenko-Pastur law. This is Theorem 2.9 from Yao et. al (2015)[2].

Theorem 1.2.1. *Let the entries $y_{i,j}$ of the $p \times n$ data matrix Y be independent random variables with mean $\mathbb{E}(y_{i,j}) = 0$ and unit variance $Var(y_{i,j}) = 1$. Let S_n be the sample covariance matrix of Y, and let $\frac{p}{n} \to c > 0$ as $(p, n) \to \infty$. Then almost surely, the empirical spectral distribution F^{S_n} converges weakly to the Marcenko-Pastur law F_c*

For example, in the figure 1.2 below you will see the density of the Marcenko-Pastur law with $c = 0.5$ and a histogram of the eigenvalues of a sample covariance matrix based on Y of size $p = 3200 \times n = 6400$ where each element $y_{i,j}$ is standard normally distributed. Because $c = 0.5$,our data lives on the interval $[a, b] = [(1 - \sqrt{c})^2, (1 + \sqrt{c})^2]$, which is $[0.0858, 2.9142]$.

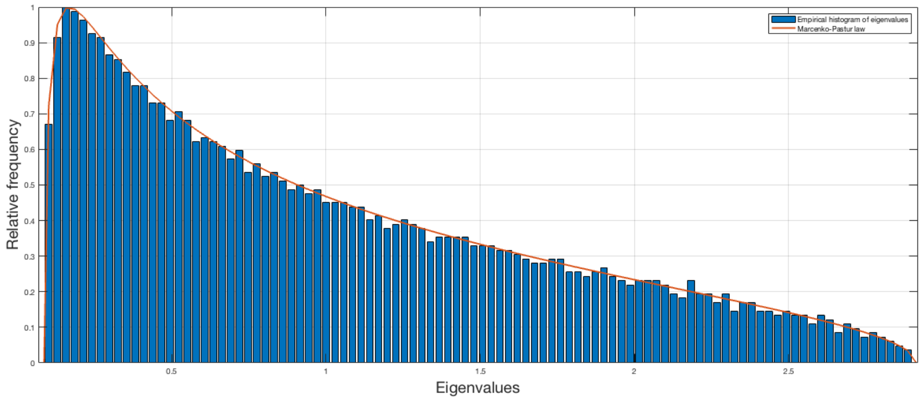


Figure 1.2: Empirical eigenvalues of a sample covariance matrix obey the Marcenko-Pastur law

Summary

- Given a $p \times p$ matrix S, with eigenvalues $\lambda_1, \ldots \lambda_p$, one computes the empirical spectral distribution function $F^S = \frac{1}{p}\sum_{i=1}^{p} \delta_{\lambda_i}$. Here δ_a denotes the Dirac mass placed at a point a.
- Eigenvalues of the sample covariance matrix S_n are distributed by some non-random law, the Marcenko-Pastur law, as $(p, n) \to \infty$

1.3 Statistics

In this paper we will also concern ourselves with statistics of the sample covariance matrix. We will make use of some tests and try to compute their sizes and powers. A common test we will would like to do is the sphericity test. This is a test that wants find out if the underlying population covariance matrix is the identity matrix:

$$H_0 : \Sigma_n = I \text{ versus } H_1 : \Sigma_n \neq I$$

This corresponds with a hypothesis that the underlying variables are all independent of each other, and have variance 1. In general to test a hypothesis we first need a statistic, computed from the sample covariance matrix, find it's distribution and then compute the sizes and powers.

Definitions

Definition 1.3.1. The *size* of a test is $\mathbb{P}(\text{reject } H_0 \mid H_0 \text{ is true})$. This is the probability of making a type 1 error, rejecting a true hypothesis. Rejecting a true hypothesis is also known as a false positive.

Definition 1.3.2. The *power* of a test is $\mathbb{P}(\text{reject } H_0 \mid H_1 \text{ is true})$. This the probability of rejecting a false hypothesis. Rejecting a false hypothesis is also known as a true positive.

Linear spectral statistics

Another central object throughout this paper will be a so called linear spectral statistic (LSS). These are functionals of the eigenvalues that have the following form:

Definition 1.3.3. Let S_n be a sample covariance matrix with eigenvalues $\lambda_1, \ldots, \lambda_p$ and with corresponding empirical distribution function $F_p(x)$. A *linear spectral statistic* $F^{S_n}(g)$ of a real function g is:

$$F^{S_n}(g) = \frac{1}{p}\sum_{i=1}^{p} g(\lambda_i) = \int_{-\infty}^{\infty} g(t)dF^{S_n}(t)$$

We can relate this to the linear algebra section, in particular Theorem 1.1.8. If g is an analytic function, the linear spectral statistic is the weighted sum of the trace of $g(A)$.

The reason why we are so interested in these linear spectral statistics is because we want to know how these behave in high-dimensional case. In particular we want to know they fluctuate around their limit, if it exists. We will see that for analytic functions, the limit exists and the fluctuations are normal.

Since our S_n is a direct product of it's eigenvalues and eigenvectors, knowledge about the transformed eigenvalues of S_n as $p \to \infty$ might allow us to construct new estimators that could be used as a better approximation than the sample covariance matrix. From Theorem 1.2.1 we have already that these eigenvalues are random, but satisfy a certain law in the high-dimensional case. Our main tool, the Central Limit Theorem 1.5.1 will help us out tremendously with these linear spectral statistics. It will tell us how linear spectral statistics will fluctuate around it's limit (if it exists).

Delta method

An important result in statistics is the Delta method, theorem 1 from [7](1932). The result states that if a random vector converges to some multivariate normal distribution in the limit, differentiable functions of that random vector are also normally distributed.

Theorem 1.3.1 (Delta method). *Let $X = (X_1, \dots, X_k)$ be random vector, and $g : \mathbb{R}^k \to \mathbb{R}^d$ be a differentiable function with derivative $\nabla g(a)$ at $a \in \mathbb{R}^k$. If we have for some $b > 0$ and $p \to \infty$:*

$$p^b\{X - a\} \xrightarrow{D} Y$$

then

$$p^b\{g(X) - g(a)\} \xrightarrow{D} [\nabla g(a)]^T Y$$

This result will be essential for finding the limiting distribution of parameters that itself are functions of normal variables. As we will see in the main result, functionals of eigenvalues are normally distributed in the limit. The Delta method then implies that estimators that make use of these functionals will also be normal.

Recall that when adding two normal variables X, Y, the resulting random variable $Z = X + Y$ is also normal and has expectation $\mathbb{E}[Z] = \mathbb{E}[X] + \mathbb{E}[Y]$, but variance that also depends on the covariance: $Var[Z] = Var[X] + Var[Y] + 2Cov(X, Y)$.

Summary

- From a dataset Y, we can compute the sample covariance matrix S_n. From this S_n, we can compute statistics such that we can produce tests and find the powers and sizes.
- If we have that a random vector X converges to some multivariate normal distribution, then for differentiable functions g, one has that $g(X)$ also converges to some multivariate normal distribution.

1.4 Complex analysis

A powerful tool from complex analysis is the Residue Theorem. This theorem turns a contour integral into a sum of residues. Before we state the theorem, we need to know a couple of things from complex analysis that appear in the residue theorem. We get our theory from the book by Freitag and Busam (2005) [4].

Definitions

Definition 1.4.1. Let γ be a closed, smooth curve that does not pass through $w \in \mathbb{C}$. The *winding number* of γ with respect to w is

$$\chi(\gamma, w) = \frac{1}{2\pi i} \oint_{\gamma} \frac{1}{z-w} dz$$

This is definition $III6.1$ from [4]. This is the rigorous definition, but the intuition is much simpler. The winding number of a curve measures how many times a curve goes around a given point. Since we will mostly concern ourselves with the a curve that is unit circle once in the counterclockwise direction, our winding number for points within the unit circle will be 1, and the winding number for all points outside the unit circle is 0.

Now we define what a residue is:

Definition 1.4.2. Let $U_r^*(c) = \{z \in \mathbb{C} : 0 < |z-c| < r\}$ be the punctured disk around $a \in \mathbb{C}$ with radius r. Let $f : U_r^*(c) \to \mathbb{C}$ be an analytic function such that c is a singularity of f, and let

$$f(z) = \sum_{n=-\infty}^{\infty} a_n (z-c)^n$$

be the Laurent series of f in $U_r^*(a)$. The *residue* $\text{Res}(f, c)$ of f at c is the coefficient a_{-1}.

This is again the formal definition $III6.2$ from Freitag, Busam (2005) [4]. However, there is a useful remark, remark $III6.4$ from [4] that gives us an easy method for computing residues. Generally, if c is a pole of f order n, then

$$Res(f, c) = \frac{1}{(n-1)!} \lim_{z \to c} \frac{d^{n-1}}{dz^{n-1}} ((z-c)^n f(z))$$

This is a handy tool to compute there residue if you know the order of the pole, which depends on what function f you have.

Theorems

Now that we know what residues and winding numbers are, we can go to the main theorem, Theorem $III6.3$ [4]

Theorem 1.4.1 (Residue Theorem)**.** *Let $D \subset \mathbb{C}$ be a simply connected open subset, that contains finite set of distinct points $\{z_1, \ldots, z_n\}$, and f an analytic function on $D \setminus \{z_1, \ldots, z_n\}$. Let γ be a closed curve in D that doesn't pass through any z_k, and let $\chi(\gamma, z_k)$ be the winding number.*

The line integral of f around γ is then equal to

$$\oint_{\gamma} f(\zeta) d\zeta = 2\pi i \sum_{k=1}^{n} \chi(\gamma, z_k) Res(f, z_k)$$

This beautiful result tells us that contour integrals of analytic functions are sums of it's individual residues. What that means for us is that the expectations and variances for linear spectral statistics can be easily computed, due to the following theorem.

1.5 Central Limit Theorem for linear spectral statistics

All of the above theory comes together in the following Central Limit Theorem from Yao, Zheng and Bai (2015) [2]. We will use it throughout the paper to find the limiting distributions of various linear spectral statistics of our sample covariance matrix in high-dimensional case.

Theorem 1.5.1 (Central limit theorem)**.** *Let the data $\{y_{i,j}\}$ of the data matrix $\boldsymbol{Y}$ be independent identically distributed random variables satisfying $\mathbb{E}(y_{i,j}) = 0, \mathbb{E}(y_{i,j}^2) = 1, \mathbb{E}(y_{i,j}^4) = \beta + 1 + \kappa < \infty$. Here $\kappa = 2$ if our data is real, and $\kappa = 1$ if our data is complex, and $\beta = \mathbb{E}(y_{i,j}^4) - 1 - \kappa$. Also assume $p \to \infty, n \to \infty, \frac{p}{n} \to c > 0$.*

Furthermore, let $f_1, \dots f_k$ be analytic functions on an open region contained in the support of F_c, which is the support of the Marcenko-Pastur law with index c. Then the random vector $\{X_n(f_1), \dots, X_n(f_k)\}$ where $X_n(f) = p\{F^{S_n}(f) - F_c(f)\}$, where $F_c(f) = \int f(x)p_c(x)dx$, converges weakly (in distribution) to a normal (Gaussian) vector $\{(X_{f_1}, \dots, X_{f_k}\}$, with mean and covariance functions:

$$\mathbb{E}(X_f) = (\kappa - 1)I_1(f) + \beta I_2(f)$$
$$Cov(X_f, X_g) = \kappa J_1(f, g) + \beta J_2(f, g)$$

where

$$I_1(f) = -\frac{1}{2\pi i}\oint \frac{c\{\underline{s}/(1+\underline{s})\}^3(z)f(z)}{[1 - c\{\underline{s}/(1+\underline{s})\}^2]^2}dz$$
$$I_2(f) = -\frac{1}{2\pi i}\oint \frac{c\{\underline{s}/(1+\underline{s}\}^3(z)f(z)}{1 - c\{\underline{s}/(1+\underline{s})\}^2}dz$$
$$J_1(f, g) = -\frac{1}{4\pi^2}\oint\oint \frac{f(z_1)g(z_2)}{(\underline{m}(z_1) - \underline{m}(z_2))^2}dz_1 dz_2$$
$$J_2(f, g) = \frac{-c}{4\pi^2}\oint f(z_1)\frac{\partial}{\partial z_1}\{\frac{\underline{s}}{1+\underline{s}}(z_1)\}dz_1 \oint g(z_2)\frac{\partial}{\partial z_2}\{\frac{\underline{s}}{1+\underline{s}}(z_2)\}dz_2$$

where the integrals are taken over contours enclosing the support of F_c.

In this current form, Theorem 1.5.1 is hard to apply directly. However there is a nice proposition, Proposition 3.6 from the book by Yao, Zheng and Bai (2015) [2], that converts the above integrals into integrals on the unit circle $|z| = 1$.

Proposition 1.5.1. *The integrals I_1, I_2, J_1 and J_2 in theorem 1.5.1 are equal to*

$$I_1(f) = \lim_{r \downarrow 1} I_1(f, r)$$
$$I_2(f) = \frac{1}{2\pi i}\oint_{|z|=1} f(|1 + hz|^2)\frac{1}{z^3}dz$$
$$J_1(f, g) = \lim_{r \downarrow 1} J_1(f, g, r)$$
$$J_2(f, g) = -\frac{1}{4\pi^2}\oint_{|z_1|=1} \frac{f(|1 + hz_1|^2)}{z_1^2}dz_1 \oint_{|z_2|=1} \frac{g(|1 + hz_2|^2)}{z_2^2}dz_2$$

with

$$I_1(f, r) = \frac{1}{2\pi i}\oint_{|z|=1} f(|1 + hz|^2)[\frac{z}{z^2 - r^{-2}} - \frac{1}{z}]dz$$
$$J_1(f, g, r) = -\frac{1}{4\pi^2}\oint_{|z_2|=1}\oint_{|z_1|=1} \frac{f(|1 + hz_1|^2)g(|1 + hz_2|^2)}{(z_1 - rz_2)^2}dz_1 dz_2$$

In these propositions, r starts close to but is greater than 1, and $h = \sqrt{c}$

Even though these integrals still look difficult, the integrals from the proposition are significantly easier to calculate than the integrals given in the central limit theorem. That is because we can directly apply the residue theorem, which is stated in the section about complex analysis. Recall that our curve is just the unit circle once, so we have winding number 1 for all points. This means we just have to calculate the individual residues at each pole and sum them together to get the contour integrals.

This theorem states that if we have a linear spectral statistic with an analytic function f, then this linear spectral statistic minus some centering term $pF_c(f)$ will be normally distributed.

1.6 Motivation and conclusions

We have seen a rich mathematical background already, but one might ask why we are so interested in all the theory about high-dimensional data analysis.

Let's say we have a given data matrix $Y = \Sigma_n^{1/2} X$, with Σ_n the true covariance matrix and X the $p \times n$ matrix whose elements are independent identically distributed with mean 0 and variance 1. Importantly, we also require the fourth moment to exist. We compute the sample covariance matrix S_n. It is understood that under $H_0 : \Sigma_n = I$, if p is fixed and we let $n \to \infty$, then S_n almost surely converges to the true $p \times p$ population covariance matrix $\Sigma_n = I$;

Since $\Sigma_n = I$, our Y is directly the noise matrix and so each element of Y on the i-th row and j-th column $y_{i,j}$, is independent with mean 0 and variance 1. Since Y was $p \times n$, we have $i \in \{1, \dots, p\}$ and $j \in \{1, \dots, n\}$. That has the following implications for the elements of S_n:

$$\begin{aligned}
\text{(On the diagonal) } \mathbb{E}[s_{i,i}] &= \mathbb{E}[\frac{1}{n}\sum_{k=1}^{n} y_{k,i}y_{k,i}] = \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}[y_{k,i}^2] = \frac{1}{n}\sum_{k=1}^{n} 1 = 1 \\
Var[s_{i,i}] &= \frac{1}{n^2}\sum_{k=1}^{n}\mathbb{E}[y_{k,i}^4] = \frac{1}{n}\mathbb{E}[y_{k,i}^4] \\
\text{(Off the diagonal) } \mathbb{E}[s_{i,j}] &= \mathbb{E}[\frac{1}{n}\sum_{k=1}^{n} y_{k,i}y_{k,j}] = \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}[y_{k,i}y_{k,j}] = 0 \text{ (Since they are independent)} \\
Var[s_{i,j}] &= \frac{1}{n^2}\sum_{k=1}^{n} Var[y_{k,i}y_{k,j}] = \frac{1}{n^2}\sum_{k=1}^{n}\mathbb{E}[y_{k,i}^2 y_{k,j}^2] \\
&= \frac{1}{n^2}\sum_{k=1}^{n}\mathbb{E}[y_{k,i}^2]\mathbb{E}[y_{k,j}^2] = \frac{1}{n^2}\sum_{k=1}^{n} 1 \cdot 1 = \frac{1}{n}
\end{aligned}$$

At the end we made use of Theorem 6.66 from Grimmett,Welsh [10], which states that for independent random variables X, Y and functions $g, h : \mathbb{R} \to \mathbb{R}$;

$$\mathbb{E}[(g(X), h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$$

In particular, we took x^2 for both g and h. In these formulas we see the need to the existence of the fourth moment for our data. This will be also be an assumption in the main theorem 1.5.1. Furthermore, we note that as $n \to \infty$, the variance of the estimators will approach zero and thus our sample covariance matrix S_n perfectly estimates $\Sigma_n = I$.
To further demonstrate, we look at the squared Frobenius norm of the difference between S_n and Σ_n:

$$||S_n - \Sigma_n||_F^2 = \sum_{i=1}^{p}\sum_{j=1}^{p} |s_{i,j} - \sigma_{i,j}|^2 \tag{1.1}$$

Here $\sigma_{i,j}$ denotes the element of Σ_n on the i-th row and j-th column, and is non-random. Using that for any random variable X and constant $a \in \mathbb{R}$, we have that

$$\mathbb{E}[|X - a|^2] = \mathbb{E}[(X - a)^2] = \mathbb{E}[X^2] - 2a\mathbb{E}[X] + a^2$$

Now, taking the expectation in expression 1.1 on both sides while applying the remark from above gives us

$$\mathbb{E}||S_n - \Sigma_n||_F^2 = \sum_{i=1}^{p}\sum_{j=1}^{p} \mathbb{E}[s_{i,j}^2] - 2\sigma_{i,j}\mathbb{E}[s_{i,j}] + \sigma_{i,j}^2$$

Let us fix $i \in \{1, \ldots, p\}$. We're left with $\sum_{j=1}^{p} \mathbb{E}[s_{i,j}^2] - 2\sigma_{i,j}\mathbb{E}[s_{i,j}] + \sigma_{i,j}^2$. We know that if $i \neq j$, the summand is equal to $\mathbb{E}[s_{i,j}^2]$, since then $\sigma_{i,j} = 0$. We then get for $i \neq j : \mathbb{E}[s_{i,j}^2] = Var[s_{i,j}] + \mathbb{E}[s_{i,j}]^2 = \frac{1}{n}$. Since we have $p-1$ elements $j \neq i$, we have in our sum $\frac{p-1}{n}$. If $i = j$, then

$$\begin{aligned}
&\mathbb{E}[s_{i,i}^2] - 2\sigma_{i,i}\mathbb{E}[s_{i,i}] + \sigma_{i,i}^2 && (1.2)\\
&\qquad = Var[s_{i,i}] + \mathbb{E}[s_{i,i}]^2 - 2\sigma_{i,i}\mathbb{E}[s_{i,i}] + \sigma_{i,i}^2 && (1.3)\\
&\qquad = \frac{1}{n}\mathbb{E}[y_{k,i}^4] + 1 - 2 \cdot 1 \cdot 1 + 1^2 && (1.4)\\
&\qquad = \frac{1}{n}\mathbb{E}[y_{k,i}^4] && (1.5)\\
& && (1.6)
\end{aligned}$$

So for any fixed $i \in \{1, \ldots, p\}$, we have that $\sum_{j=1}^{p} |s_{i,j} - \sigma_{i,j}|^2 = \frac{1}{n}\mathbb{E}[y_{k,i}^4] + \frac{p-1}{n}$. Now taking the sum over all $i = 1 \ldots p$, we get that

$$\mathbb{E}||S_n - \Sigma_n||_F^2 = p\{\frac{1}{n}\mathbb{E}[y_{k,i}^4] + \frac{p-1}{n}\} = \frac{p}{n}\mathbb{E}[y_{k,i}^4] + \frac{p(p-1)}{n}$$

Notice that when we fix p, but let $n \to \infty$, the expectation converges almost surely to 0, and thus S_n will be close to $\Sigma_n = I$ in the limit. We still require the existence of the fourth moment of the data.

Now we want to turn our attention to the high-dimensional case $p \to \infty$, with $\frac{p}{n} \to c > 0$ In these case things are a bit different. We again get

$$\mathbb{E}||S_n - \Sigma_n||_F^2 = \frac{p}{n}\mathbb{E}[y_{k,i}^4] + \frac{p(p-1)}{n}$$

However, this will no longer converge to 0, but diverge to $c \cdot \mathbb{E}[y_{k,i}^4] + c(p-1) = \infty$.

This is however not really a fair comparison though, since $S_n - \Sigma_n$ is a $p \times p$ matrix and therefore an infinite matrix so one could reasonably expect that the sum over it's elements squared will diverge. But now let's take a look at the normalized expectation of the squared Frobenius norm:

$$\frac{1}{p}\{\mathbb{E}||S_n - \Sigma_n||_F^2\} = \frac{1}{n}\mathbb{E}[y_{k,i}^4] + \frac{(p-1)}{n}$$

Now taking the limit as $(p, n) \to \infty$, $\frac{p}{n} \to c > 0$:

$$\lim_{(p,n)\to\infty} \frac{1}{n}\mathbb{E}[y_{k,i}^4] + \frac{(p-1)}{n} = 0 + c = c$$

So even the normalized expectation won't converge to 0, but to the ratio $\frac{p}{n} = c$. Clearly there is some non-negligible discrepancy in the high-dimensional case.

We can also look at this from a perspective using the eigenvalues. In the case of a fixed dimension p, the eigenvalues of S_n all converge to the eigenvalues of Σ_n. If $H_0 : \Sigma_n = I$ holds, then all p eigenvalues will converge almost surely to 1:

Theorem 1.6.1. *Let p be fixed and S_n be the sample covariance matrix, and assume $S_n \to I$ as $n \to \infty$, elementwise. Then the eigenvalues of S_n will converge to 1.*

Proof. Let S_n be a sample covariance matrix and $\lim_{n\to\infty} S_n = I$. Let $\lambda_{i,n}, i \in \{1, \ldots, p\}$ be the eigenvalues of S_n. For any λ_i we have that by definition, with non-zero eigenvector v_i

$$\begin{aligned}
S_n v_i &= \lambda_{i,n} v_i\\
S_n v_i - v_i &= \lambda_{i,n} v_i - v_i\\
(S_n - I)v_i &= (\lambda_{i,n} - 1)v_i\\
\lim_{n\to\infty} (S_n - I)v_i &= \lim_{n\to\infty} (\lambda_{i,n} - 1)v_i\\
\mathbf{0} &= \lim_{n\to\infty} (\lambda_{i,n} - 1)v_i
\end{aligned}$$

Now v_i was nonzero, so we must have that $\lim_{n\to\infty}(\lambda_{i,n} - 1) = 0$. Since $\lambda_{i,n}$ was any eigenvalue, all eigenvalues converge to 1. □

However in the high-dimensional case, we have found by Theorem 1.2.1 that the eigenvalues are spread out over the interval $[a, b] = [(1 \mp \sqrt{c}^2]$. So in the case where $p \to \infty$ as well, the eigenvalues of the sample covariance matrix will not converge to 1, but instead they obey the MP-law. So the eigenvalues of S_n will not be consistent estimators for the true eigenvalues. Using the sample covariance matrix directly in high-dimensional statistics might lead to serious errors.

Concluding say we have the following situation:

- The data matrix $Y_n = \Sigma_n^{1/2} X_n$, with each $x_{i,j}$ iid with mean 0 and variance 1, with existing fourth moment
- The sample covariance matrix S_n with eigenvalues $\lambda_1 \dots \lambda_p$
- An analytic function $g : \mathbb{R} \to \mathbb{R}$

Then by Theorem 1.2, the eigenvalues obey the Marcenko-Pastur law if $(p, n) \to \infty$. Consequently the linear spectral statistics $F^{S_n}(g)$ converges to $\int g(x)dF_c(x)$, with F_c the cumulative distribution function of the MP-law with index c. Moreover, by the Central Limit Theorem 1.5.1, the fluctuations of the linear spectral statistic around it's limit are normal:

$$\begin{aligned} & p\{\frac{1}{p}\sum_{i=1}^{p} g(\lambda_i) - \int g(x)dF_c\} \\ &= p\{\int g(x)dF^{S_n}(x) - \int g(x)dF_c(x)\} \\ &= p\int g(x)d(F^{S_n}(x) - F_c(x)) \xrightarrow{D} N(\mu_g, \sigma_g^2) \end{aligned}$$

Formulae for μ_g and σ_g^2 are given in the Central Limit Theorem. Using this machinery, we will try to find new test statistics based on new types of estimators.

Chapter 2

Sphericity tests in high-dimensional case

As stated in section 1.3, one test we would like to perform in the high-dimensional case is the so called *sphericity* test. This is a test in which we want to test the following hypotheses:

$$H_0 : \Sigma_n = \sigma^2 I \text{ versus } H_1 : \Sigma_n \neq \sigma^2 I$$

We want to find out if our true underlying covariance matrix is a multiple of the identity matrix. This corresponds with testing if the underlying variables of our population are independent, and have variance σ^2. As we will see in a minute, the statistics that will be of interested are independent of σ, so we can take without loss of generality that $\sigma^2 = 1$, so our test reduces to:

$$H_0 : \Sigma_n = I \text{ versus } H_1 : \Sigma_n \neq I$$

Some mathematicians have found a way to make tests from finite dimensional case also work in high-dimensional case. In this section I will discuss two of these tests for the high-dimensional case, the corrected likelihood ratio test (CLRT) and the corrected John's test (CJ). Before I can go deeper, we need to make some assumptions on our data: [1]

- Our observed $p \times n$ data matrix Y is the product of a true positive definite $p \times p$ covariance matrix Σ_n and a $p \times n$ matrix X which elements $x_{i,j}$ are independent identical random variables with mean 0 and variance 1. Only $Y = \Sigma_n^{1/2} X$ is observed.
- Let $H_n(t)$ be the distribution function of the eigenvalues of Σ_n. Then we assume that $H_n(t)$ converges to some limit $H(t)$ as $n \to \infty$ at all points of continuity of $H(t)$

Throughout the paper we will use an indicator κ just like in the Central Limit Theorem 1.5.1, which is 2 if our data is real and 1 if our data is complex. Most of our simulation however will assume standard normal data or $\Gamma(4, 2) - 2$ data. Subtracting 2 from the $\Gamma(4, 2)$ variable makes it that the expectation $\mathbb{E}[\Gamma(4, 2) - 2]$, where 4 is the shape and 2 is the rate, is equal to $\frac{4}{2} - 2 = 0$. The variance is equal to $Var[\Gamma(4, 2)] = \frac{4}{2^2} = 1$. Subtracting 2 won't change the variance, so we have that $\Gamma(4, 2) - 2$ has mean 0 and variance 1, and so we have another distribution to simulate from that satisfies the null hypothesis. We will leave κ in the derivations however, to derive the theory as general as possible.

We also had a parameter $\beta = \mathbb{E}[x_{i,j}^4] - \kappa - 1$ in the Central Limit Theorem, a parameter that depends on the fourth moment. In the case of standard normal data, one readily checks that the 4th moment is $\mathbb{E}[x_{i,j}^4] = 3$, and for $\Gamma(4, 2) - 2$, the 4th moment is 4.5. Then $\beta = 3 - 1 - 2 = 0$ or $\beta = 4.5 - 1 - 2 = 1.5$ respectively. We will leave β also in the derivation, but we will substitute it for it's appropriate value when simulating.

2.1 Corrected Likelihood Ratio Test (CLRT)

For finite dimensional statistics, a commonly used test is the regular likelihood ratio test. As a matter of fact, by the Neyman-Pearson lemma it is the most powerful test. It is then interesting to study what the test would looks like and behaves in the high-dimensional case.

The likelihood ratio test statistic for a sample covariance matrix S_n with eigenvalues $\lambda_1, \dots \lambda_p$ is [8]

$$L_n = \Big(\frac{(\lambda_1 \cdots \lambda_p)^{(1/p)}}{\frac{1}{p}(\lambda_1 + \cdots + \lambda_p)} \Big)^{(\frac{1}{2}pn)}$$

The correction for the high-dimensional case is from the following Theorem 2.1, from the paper by Wang, Yao (2013)[3]:

Theorem 2.1.1. *Let $\Lambda_n = -\frac{2}{n}\log(L_n)$ be the test statistic, with L_n from above. Assume our data $x_{i,j}$ are independent identically distributed, satisfying $\mathbb{E}[x_{i,j}] = 0, \mathbb{E}[x_{i,j}^2] = 1, \mathbb{E}[x_{i,j}^4] < \infty$. Then under $H_0 : \Sigma_n = I$:*

$$\Lambda_n + (p-n) \cdot \log(1 - \frac{p}{n}) - p \xrightarrow{D} N\big\{ -\frac{\kappa - 1}{2}\log(1-c) + \frac{1}{2}\beta c, -\kappa\log(1-c) - \kappa c \big\}$$

This is the corrected likelihood ratio test statistic in the high-dimensional case under the null hypothesis. We will afterwards refer to this as the CLRT. One notices that this statistic is independent of σ and that the statistic depends on the logarithm of $1 - c$ in both the expectation and the variance. In particular, if c is close to 1, so p is close to n, then the variance will blow out of proportion so one could expect the power of this test to break down if p is large enough compared to n. Nevertheless, we would like to study this test and compare it to other statistics to find out if the optimality of power caries over to high-dimensional statistics.

2.2 Corrected John's test (CJ)

Another test that has been altered to work in the high dimensional case is the corrected John's test, or CJ test, from the paper by John [9]. This test is the most powerful test which is rotation invariant. That means if you rotate the noise matrix X by some orthogonal matrix Q, then the John's statistic computed from the new sample covariance matrix

$$S_n = \frac{1}{n}YY^T = \frac{1}{n}\Sigma^{1/2}QXX^TQ^T(\Sigma^{1/2})^T$$

is the most powerful among all alternatives. We will not rotate our data in this paper, but we note that the identity matrix is also sort of a rotation matrix. If we take $Q = I$, then still the John's test should be really powerful.

In the finite dimensional case he proposed to use a statistic of the form

$$T_2 = \frac{p^2 n}{2} tr\big\{ \frac{S_n}{tr(S_n)} - \frac{I}{p} \big\}^2$$

When p is fixed and $n \to \infty$, it can be shown that under the null hypothesis this statistic $T_2 \xrightarrow{D} \chi^2_f$, a Chi-squared distribution with $f = \frac{1}{2}p(p+1) - 1$ degrees of freedom.

The *Corrected John's* test in the high-dimensional case is constructed a bit differently. We now define a U-statistic, which uses the T_2 statistic from above: $U = \frac{2}{pn}T_2$. We then get the following theorem [3]

Theorem 2.2.1. *Let $x_{i,j}$ be independent identically distributed random variables, satisfying $\mathbb{E}[x_{i,j}] = 0, \mathbb{E}[x_{i,j}^2] = 1, \mathbb{E}[x_{i,j}^4] < \infty$. Let $U = \frac{2}{pn}T_2$ be the test statistic, computed from S_n. Then under $H_0 : \Sigma_n = I$ and $\frac{p}{n} \to c > 0$:*

$$nU - p \xrightarrow{D} N(\kappa - 1 + \beta, 2\kappa)$$

As said, a motivation for considering the CJ test is that this is a test that is optimized for power already in the finite dimensional case under rotation invariance. [9] We would like to find out if the power carries over to the high-dimensional case. What is important to note, is that the limiting distribution is independent from the ratio $c = \frac{p}{n}$, unlike the CLRT test.

Chapter 3

Linear shrinkage estimators

In this chapter we will derive a new test for the high-dimensional case based on a different estimator, namely the linear shrinkage estimator, based on the work done by Bodnar et.al (2014)[1]. We assume the same setting as for the CLRT and CJ test, i.e. we only observe $Y = \Sigma_n^{1/2}X$, with each $x_{i,j}$ independent with mean 0 and variance 1, and the distribution function $H_n(t)$ converges to some limit $H(t)$.

The general linear shrinkage estimator Σ_{GLSE} is of the form:

$$\Sigma_{GLSE} = \alpha_n S_n + \beta_n \Sigma_0 \tag{3.1}$$

Here Σ_0 is a symmetric positive definite matrix bounded in trace norm, i.e $\exists N > 0 : ||\Sigma_0||_{tr} = tr(\Sigma_0) \leq N$. It could be interpreted as a prior belief of what our true Σ_n could be. The reason why we require Σ_0 to be a symmetric positive definite matrix is that any symmetric positive definite matrix is a covariance matrix for some multivariate distribution.

For example, let M be any positive definite matrix. From our matrix theorem 1.1.4 all eigenvalues are real and positive. By Theorem 1.1.2, we can decompose our matrix M into $C^{-1}DC$, where in particular D consists of the eigenvalues. Since all eigenvalues are positive, we can take the real square root, to have $M = C^{-1}U^2C$, where $U = D^{1/2}$. Now we have that $M^{1/2} = C^{-1}UC$. Now if we consider a random vector X with mean 0 and covariance matrix I, then the random vector $Y = M^{1/2}X$ has covariance matrix M.

So, as long as we take a positive definite matrix as a prior belief, it is automatically a covariance matrix. It is of course logical to only the consider the prior belief matrices that are covariance matrices.

Now that we know the form of our estimator, we want to find the optimal shrinkage estimators for α_n and β_n.

3.1 Optimal linear shrinkage estimators

We define a quantity

$$L_f^2 = ||\Sigma_{GLSE} - \Sigma_n||_F^2 \tag{3.2}$$

This quantity measures how "far" in squared Frobenius norm our GLSE is from the true Σ_n. The smaller it is, the better Σ_n is approximated by our GLSE matrix Σ_{GLSE}. If we insert expression 3.1 into the quantity 3.2 we get:

$$\begin{aligned} L_f^2 &= ||\Sigma_{GLSE} - \Sigma_n||_F^2 \\ &= ||\alpha_n S_n + \beta_n \Sigma_0 - \Sigma_n||_F^2 \end{aligned}$$

As done in [1], working this expression out, taking partial derivatives with respect to α_n and β_n and setting those to 0, we get a system of linear equations:

$$\frac{\partial L_f^2}{\partial \alpha_n} = \alpha_n ||S_n||_F^2 + \beta_n tr(S_n \Sigma_0) - tr(S_n \Sigma_n) = 0,$$
$$\frac{\partial L_f^2}{\partial \beta_n} = \alpha_n tr(S_n \Sigma_0) + \beta_n ||\Sigma_0||_F^2 - tr(\Sigma_n \Sigma_0) = 0$$

From these equations you can find optimal parameters:

$$\alpha_n^* = \frac{tr(S_n\Sigma_n)||\Sigma_0||_F^2 - tr(\Sigma_n\Sigma_0)tr(S_n\Sigma_0)}{||S_n||_F^2||\Sigma_0||_F^2 - (tr(S_n\Sigma_0))^2}$$
$$\beta_n^* = \frac{tr(\Sigma_n\Sigma_0)||S_n||_F^2 - tr(S_n\Sigma_n)tr(S_n\Sigma_0)}{||S_n||_F^2||\Sigma_0||_F^2 - (tr(S_n\Sigma_0))^2}$$

These parameters are not usable just yet, since it depends on the unobservable Σ_n. However, it can be shown that these parameters converge to some non-random quantity in the limit.

Remark. *Two sequences p_n and q_n are asymptotically equivalent if as $n \to \infty$:*

$$|p_n - q_n| \to 0$$

Proposition 3.1.1. *Let $\frac{p}{n} \to c > 0$, and $n, p \to \infty$. The optimal shrinkage intensities then converge almost surely to some asymptotic optimal values:*

$$|\alpha_n^* - \alpha^*| \xrightarrow{a.s.} 0$$
$$|\beta_n^* - \beta^*| \xrightarrow{a.s.} 0$$

These asymptotic optimal values satisfy:

$$\alpha^* = 1 - \frac{\frac{c}{p}||\Sigma_n||_{tr}^2||\Sigma_0||_F^2}{(||\Sigma_n||_F^2 + \frac{c}{p}||\Sigma_n||_{tr}^2)||\Sigma_0||_F^2 - (tr(\Sigma_n\Sigma_0))^2}$$
$$\beta^* = \frac{tr(\Sigma_n\Sigma_0)}{||\Sigma_0||_F^2}(1 - \alpha^*)$$

These parameters are nonrandom, but still depend on the underlying true covariance matrix and are hence unknown. However, we are able to find consistent estimators for the asymptotic optimal value. [1]

$$\hat{\alpha}^* = 1 - \frac{\frac{1}{n}tr(S_n)^2||\Sigma_0||_F^2}{||S_n||_F^2||\Sigma_0||_F^2 - (tr(S_n\Sigma_0))^2}$$
$$\hat{\beta}^* = \frac{tr(S_n\Sigma_0)}{||\Sigma_0||_F^2}(1 - \hat{\alpha}^*)$$

These are the quantities that we were after. These parameters will consistently estimate the optimal asymptotic value, the value that minimizes the difference in Frobenius norm between our Σ_{GLSE} and the true covariance matrix Σ_n in the limit $(p, n) \to \infty$. Note that these values are still dependent on our choice for Σ_0.

We would now like to establish what the limits of these parameters are as $(p, n) \to \infty$. We will however only concern ourselves with $\hat{\alpha}^*$, because we observe that $\hat{\beta}^*$ itself is a function of $\hat{\alpha}^*$.

We see that $\hat{\alpha}^*$ is a function of the squared Frobenius norm $||S_n||_F^2$ and the trace $tr(S_n)$, and puts out a single real number. This observation about the structure of $\hat{\alpha}^*$ gives good motivation for the Delta method, Theorem 1.3.1.

If we want to apply the Delta method however, we first need to find the asymptotic distributions of the squared Frobenius norm $||S_n||_F^2$ and the trace $tr(S_n)$. Here we recall that the squared Frobenius norm for symmetric matrices S_n was equal to $tr(S_n^2)$. By theorem 1.1.8, with $f(x) = x^2$, we can turn the squared Frobenius norm into a linear spectral statistic. By the Central Limit Theorem, 1.5.1, this quantity minus some centering term is normally distributed in the limit $(p, n) \to \infty$. The same applies to the trace $tr(S_n)$, with $f(x) = x$. The squared Frobenius norm and trace are then perfectly applicable for the Delta method. We however need to find the limiting distribution of the squared Frobenius norm and the trace.

In the next section we state the limiting behaviour of these statistics under the null hypothesis $H_0 : \Sigma_n = I$, their derivation can be found in the appendix. The requirement that $\Sigma_n = I$ is needed to find the distributions, since the Central Limit Theorem 1.5.1 we stated only applies to independent identical random variables with mean 0 and variance 1.

3.1.1 Prior belief Σ_0

We need to state our prior belief for Σ_0, since our optimal $\hat{\alpha}$ also depends on this matrix. We will restrict to a prior belief that $\Sigma_0 = I$ as well. However, we have to keep in mind that when we define a prior belief, our linear shrinkage estimators will optimize with regard to this prior belief. If we choose another Σ_0 as a prior belief, the distribution of our optimal $\hat{\alpha}$ will change.

3.2 Limiting behaviour of $||S_n||_F^2$ and $tr(S_n)$

$||S_n||_F^2$

We note that $||S_n||_F^2 = tr(S_n^2)$. Now we apply our matrix theory. First we use the fact that sample covariance matrix S_n is real symmetric. It is real because the data are real numbers, and it's a symmetric matrix because the covariance between two estimators is a symmetric relationship. This can be seen as follows:

$$S_n = \frac{1}{n} YY^T \implies (S_n)^T = \frac{1}{n}(YY^T)^T = \frac{1}{n} YY^T = S_n$$

Because S_n is a real symmetric matrix, it is diagonalizable by Theorem 1.1.1. And because it is diagonalizable, we have from Theorem 1.1.8 that $tr(S_n^2) = \sum_{i=1}^p \lambda_i^2$. So to summarize:

$$||S_n||_F^2 = tr(S_n^2) = \sum_{i=1}^p \lambda_i^2$$

This is a linear spectral statistic of S_n with function $f = x^2$. Now we can apply the Central Limit Theorem under $H_0 = \Sigma_n = I$. It can then be shown that:

$$||S_n||_F^2 - p(1 + c)$$

is normal as $p \to \infty$ with

$$\mathbb{E}[||S_n||_F^2 - p(1 + c)] = (\kappa - 1 + \beta)c$$
$$Var[||S_n||_F^2 - p(1 + c)] = 4(\kappa + \beta)(c^3 + 2c^2 + c) + 2\kappa c^2.$$

The full derivation can be found in the appendix.

$tr(S_n)$

Similar to the squared Frobenius norm, this is just a linear spectral statistic $\sum_{i=1}^p \lambda_i$. We again use the Central Limit Theorem with $f = x$, to obtain that under the null hypothesis $H_0 : \Sigma_n = I$, we have that $tr(S_n) - p$ is normal in the limit $p \to \infty$ with

$$\mathbb{E}[tr(S_n) - p] = 0$$
$$Var[tr(S_n) - p] = (\kappa + \beta)c$$

The full derivation can be found in the appendix.

Covariance of $tr(S_n)$ and $||S_n||_F^2$

As we will see later on we will need to find the covariance between $tr(S_n)$ and $||S||_F^2$. Luckily, the Central Limit Theorem 1.5.1 and Proposition 1.5.1 give us a formula for the covariance. The derivation can again be found in the appendix. One finds for the covariance

$$Cov(tr(S_n), ||S_n||_F^2) = 2(\kappa + \beta)(c^2 + c)$$

After I had already done the derivations for the distributions and the covariance for myself, I found out that this work was already done by Wang and Yao, also in their paper from 2015 [3]. It may be good to say this to prevent any misunderstandings. On the bright side, it reassured me because I had done the derivations correctly.

3.3 Limiting distribution of optimal shrinkage intensity $\hat{\alpha}^*$

Recall from earlier that the consistent optimal estimators are:

$$\hat{\alpha}^* = 1 - \frac{\frac{1}{n}||S_n||_{tr}^2 ||\Sigma_0||_F^2}{||S_n||_F^2 ||\Sigma_0||_F^2 - (tr(S_n\Sigma_0))^2}$$
$$\hat{\beta}^* = \frac{tr(S_n\Sigma_0)}{||\Sigma_0||_F^2}(1 - \hat{\alpha}^*)$$

For the squared Frobenius norm of $\Sigma_0 = I$ we get

$$||\Sigma_0||_F^2 = tr(\Sigma_0^2) = tr(I^2) = p$$

Now we substitute $\Sigma_0 = I$ with the corresponding values into our estimators. Those estimators become

$$\hat{\alpha}^* = 1 - \frac{\frac{p}{n}||S_n||_{tr}^2}{p||S_n||_F^2 - (tr(S_n))^2} \qquad = 1 - \frac{c \cdot tr(S_n)^2}{p||S_n||_F^2 - (tr(S_n))^2}$$

Now these parameters are in a form such that we can apply the Delta method. Because the Delta method turns the distribution into a sum of two normal variables, you also need the covariance between $tr(S_n)$ and $||S_n||_F^2$. I will just state the results here, the derivation is found in the appendix again in section A.2.

Theorem 3.3.1. *Let X be a $p \times n$ data matrix consisting of independent identically distributed random variables with mean 0 and variance 1, and let $Y = \Sigma_n^{1/2} X$. Let $S_n = \frac{1}{n}YY^T$ be it's sample covariance matrix with eigenvalues $\lambda_1 \dots \lambda_n$. Let $(p, n) \to \infty$ and $\frac{p}{n} \to c > 0$. We then have that under H_0:*

$$p\hat{\alpha}^* \xrightarrow{D} N(\kappa - 1 + \beta, 2\kappa)$$

Where $\hat{\alpha}^*$ is the optimal consistent estimator of the form

$$\hat{\alpha}^* = 1 - \frac{c \cdot tr(S_n)^2}{p||S_n||_F^2 - (tr(S_n))^2}$$

Proof. The full proof can be found in appendix, section A.2. A small outline is as follows: We have

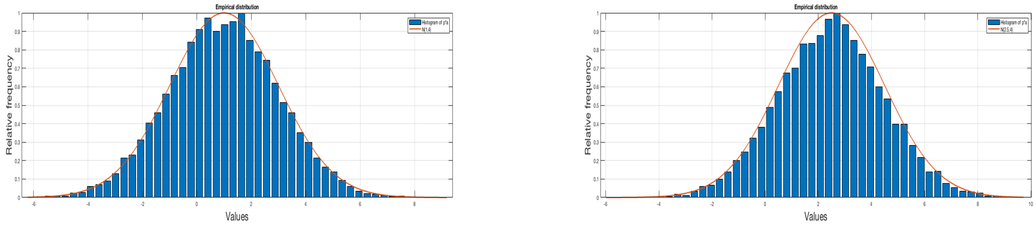
$$p\Big\{ \begin{bmatrix} \frac{1}{p}||S||_F^2 - (1+c) \\ \frac{1}{p}tr(S) - 1 \end{bmatrix} \Big\} \xrightarrow{D} \begin{bmatrix} \mathbf{X_1} \\ \mathbf{X_2} \end{bmatrix}$$

Then directly applying the Delta-method, we have that

$$p\Big\{ g(\frac{1}{p}||S||_F^2, \frac{1}{p}tr(S)^T) - g([1+c, 1]^T) \Big\} \xrightarrow{D} \nabla g([1+c, 1]^T) \begin{bmatrix} \mathbf{X_1} \\ \mathbf{X_2} \end{bmatrix} \sim N(\kappa - 1 + \beta, 2\kappa)$$

□

In figure 3.1a and 3.1b you can see the empirical distribution function of $p\hat{\alpha}^*$, repeated 10000 times with standard normally distributed or $Gamma(4,2) - 2$ data, where we have $p = 320, n = 640$. These parameters were chosen large to demonstrate the limiting behaviour better. In the case of standard normal data, we have that it should be approximately a $N(1,4)$ distribution, and for the $Gamma(4,2) - 2$ data, it should be $N(2.5,4)$



(a) Empirical distribution for standard normal

(b) Empirical distribution for $Gamma(4,2) - 2$

We see that in both cases the distributions are normal. More importantly, this is a ready to use statistic in the high-dimensional case that makes use of a different type of estimator than the regular sample covariance matrix S_n. In chapter 4 we will compare this statistic to the existing CLRT and CJ test from chapter 2.

One thing to note however is that is possible that the optimal estimator $\hat{\alpha}^*$ is smaller than 0, since $p \cdot \hat{\alpha}^* \sim N(\kappa - 1 + \beta, 2\kappa)$ and $p > 0$. This is a bit odd intuitively, but keep in mind that this is an optimization problem involving a random matrix. It is possible that the optimal solution would require a negative $\hat{\alpha}^*$. However this only happens in the finite case. If you take the limit as $(p,n) \to \infty$ one would never find a negative optimal estimator $\hat{\alpha}^*$.

3.4 Equivalence between $H_0 : \Sigma_n = I$ and $H_0 : \alpha^* = 0$

I would like to point a relation between the null hypothesis $H_0 : \Sigma_n = I$ and optimal shrinkage intensity $\alpha^* = 0$. Recall it's expression from proposition 3.1.1:

$$\alpha^* = 1 - \frac{\frac{c}{p}||\Sigma_n||_{tr}^2 ||\Sigma_0||_F^2}{(||\Sigma_n||_F^2 + \frac{c}{p}||\Sigma_n||_{tr}^2)||\Sigma_0||_F^2 - (tr(\Sigma_n \Sigma_0))^2}$$

Plugging in our null hypothesis $H_0 : \Sigma_n = I$ and our prior belief $\Sigma_0 = I$ we get

$$\begin{aligned}\alpha^* &= 1 - \frac{\frac{c}{p}||\Sigma_n||_{tr}^2 ||\Sigma_0||_F^2}{(||\Sigma_n||_F^2 + \frac{c}{p}||\Sigma_n||_{tr}^2)||\Sigma_0||_F^2 - (tr(\Sigma_n \Sigma_0))^2} \\ &= 1 - \frac{\frac{c}{p}p^2 \cdot p}{(p + \frac{c}{p}p^2)p - p^2} \\ &= 1 - \frac{cp^2}{p^2 + cp^2 - p^2} = 1 - 1 = 0\end{aligned}$$

The optimal shrinkage estimator $\hat{\alpha}^*$ is asymptotically equivalent with $\alpha^* = 0$. This can be interpreted as follows: If our prior belief Σ_0 perfectly corresponds with the true covariance matrix, then the optimal linear shrinkage estimator $\Sigma_{GLSE} = \alpha_n S_n + \beta_n \Sigma_0$ will be equal to $\beta^* \Sigma_0$ and not depend at all on the sample covariance matrix in the limit. The closer your initial prior belief is to the true covariance matrix, the smaller the optimal α^* will be.

This agrees with was said in subsection 3.1.1. There we saw that our optimal estimator $\hat{\alpha}^*$ will depend on what we define as a prior belief. If we have a prior belief Σ_0 that is not at all close to the true covariance matrix, then the sample covariance matrix will play a larger role and thus one expects the optimal estimator $\hat{\alpha}^*$ not to converge to 0 at all, but to some other value.

Now conversely, let's assume $\alpha^* = 0$. We get for the optimal linear shrinkage estimator, with again prior belief $\Sigma_0 = I$

$$\begin{aligned}\Sigma_{OLSE} &= \alpha^* I + \beta^* \Sigma_0 \\ &= \alpha^* S_n + \frac{tr(\Sigma_n \Sigma_0)}{||\Sigma_0||_F^2}(1-\alpha^*)\Sigma_0 \\ &= \frac{tr(\Sigma_n)}{p} I\end{aligned}$$

Recall that we assumed our Σ_{OLSE} to be optimal, and so it minimizes the Frobenius norm. By inspection it is not hard to see that the only matrix that minimizes the squared Frobenius norm is a multiple of the $p \times p$ identity and the squared norm is 0:

$$\begin{aligned}||\Sigma_{OLSE} - \Sigma_n||_F^2 &= ||\frac{tr(a\Sigma_n)}{p} I - a\Sigma_n||_F^2 \\ = ||\frac{ap}{p} I - aI||_F^2 &= ||aI - aI||_F^2 = 0\end{aligned}$$

If we had any other matrix than a multiple of the identity as a solution, it would have at least non-zero element $\sigma_{i,j}$ somewhere. It then contributes $\sigma_{i,j}^2$ to the Frobenius norm, and thus would be greater than 0. However since we assumed our data to have variance 1 and not any $a \in \mathbb{R}$, the only matrix that satisfies both requirements is just the identity. Thus we have $\alpha^* = 0 \implies \Sigma_n = I$.

The equivalence is thus as follows:

$$H_0 = \Sigma_n \iff \alpha^* = 0, \text{ given that } \Sigma_0 = I$$

Testing for $\alpha^* = 0$ is then just equivalent with testing the null hypothesis $H_0 : \Sigma_n = I$

3.5 Relation between CJ and linear shrinkage

In the previous section we have derived a new statistic in the high-dimensional case, namely

$$p\hat{\alpha}^* = p \times \{1 - \frac{c \cdot tr(S_n)^2}{p||S_n||_F^2 - (tr(S_n))^2}\} \xrightarrow{D} N(\kappa - 1 + \beta, 2\kappa), \text{ as } (p,n) \to \infty$$

We immediately see that this statistic has the same asymptotic distribution as the CJ test from section 2.2. This is a quite striking observation, since the statistics $nU - p$ and $p\hat{\alpha}^*$ are in general not equal to each other;

We can rewrite $nU - p$ in the following way:

$$\begin{aligned}nU - p &= n\frac{2}{pn}T_2 - p \\ &= n\frac{2}{pn}\frac{p^2 n}{2} tr((\frac{S_n}{tr(S_n)} - \frac{I}{p})^2) - p \\ &= np \cdot tr(\frac{S^2}{tr(S)^2} - 2\frac{S_n}{ptr(S_n)} + \frac{I^2}{p^2}) - p \\ &= np\frac{tr(S_n^2)}{tr(S_n)^2} - 2n + n - p \\ &= np\frac{tr(S_n^2)}{tr(S_n)^2} - n - p = np\frac{||S_n||_F^2}{tr(S_n)^2} - n - p\end{aligned}$$

This is in general not equal to $p\hat{\alpha}^* = p\{1 - \frac{ctr(S_n^2}{p||S_n||_F^2 - tr(S_n)^2}\}$. As a matter of fact, these seem not very related at all, except that they both depend on the squared Frobenius norm and trace in some

way, but this is not really unique.
One object we could then maybe study is the absolute difference between $nU - p$ and $p\hat{\alpha}^*$.

$$|nU - p - p\hat{\alpha}^*|$$

Because they have the same asymptotic distribution, their difference will center around 0. However it's possible that, when we multiply by p or n, that the difference becomes normal. For example, we already saw $\hat{\alpha}^*$ was asymptotically 0, but when multiplied by p, the quantity $p\hat{\alpha}^*$ was normally distributed.

Since they are equally distributed, tests that make us of either the CJ statistic or the LS statistic $p\hat{\alpha}^*$ will behave the same in high-dimensional case. However, we cannot test with infinite p or n, and it could be that one outperforms the other at an earlier point, that is; is there a p where one of the test already shows high-dimensional behaviour where the other does not. We would like to find if there are any significant differences between these two tests when simulating.

Chapter 4

Simulation study linear shrinkage

In this section we will test the new linear shrinkage statistic qualitatively in comparison with the existing corrected tests from section 2. We will do that in the following matter: We try to find empirical powers as the *distance* from the alternative $H_1 : \Sigma_n \neq I$ to the null hypothesis $H_0 : \Sigma_n = I$ increases. We take a look at three ways of increasing the distance. How that is done precisely we will discuss shortly.

Setting

Again, like in section 2, we have the following setting for our data:

- Our observed $p \times n$ data matrix Y is the product of a true positive definite $p \times p$ covariance matrix Σ_n and a $p \times n$ matrix X which elements $x_{i,j}$ are *real* independent identical random variables with mean 0 and variance 1. Only $Y = \Sigma_n^{1/2} X$ is observed.
- Let $H_n(t)$ be the empirical distribution function of the eigenvalues of Σ_n. Then we assume that $H_n(t)$ converges to some limit $H(t)$ at all points of continuity of $H(t)$

We will assume as our null hypothesis $H_0 : \Sigma_n = I$. This implies that it's eigenvalues are all 1, so the above $H_n(t)$ is δ_1, a Dirac mass at 1, for all n. So the second point is satisfied. Because $\Sigma_n = I$, Y is directly equal to X, and thus has i.i.d. elements, with mean 0 and variance 1. Our sample covariance matrix S_n is computed as $S_n = \frac{1}{n} Y Y^T$. From that we compute the 3 statistics of interest with limiting behaviour $(p, n) \to \infty$, $\frac{p}{n} \to c \in (0, 1)$:

$$\begin{aligned} \Lambda_n - (p-n)\log(1-c) - p &\xrightarrow{D} N\big\{ -\frac{\kappa - 1}{2}\log(1-c) + \frac{1}{2}\beta c, -\kappa \log(1-c) - \kappa c\big\} \\ nU - p &\xrightarrow{D} N(\kappa - 1 + \beta, 2\kappa) \\ p\hat{\alpha}^* &\xrightarrow{D} N(\kappa - 1 + \beta, 2\kappa) \end{aligned}$$

We will mostly be working with $x_{i,j}$ which are real standard normal random variables, we can substitute $\kappa = 2$, and as a result $\beta = 0$, because the 4th moment of standard normal variables is 3. That turns the distributions into:

$$\begin{aligned} \Lambda_n - (p-n)\log(1-c) - p &\xrightarrow{D} N\big\{ -\frac{1}{2}\log(1-c), -2\log(1-c) - 2c\big\} \\ nU - p &\xrightarrow{D} N(1, 4) \\ p\hat{\alpha}^* &\xrightarrow{D} N(1, 4) \end{aligned}$$

We now want to compare these statistics with the following test:

$$H_0 : \Sigma_n = I \text{ versus } H_1 : \Sigma_n \neq I$$

We have to define a rejection condition. For that we need a significance level α. This is chosen as the probability of rejecting a true null hypothesis. We reject the H_0 if our statistics are too far from

their limiting distribution. For a general statistic with known distribution $T \sim N(\mu, \sigma^2)$, we reject the null hypothesis if $\frac{T-\mu}{\sigma} > w(\alpha)$. The value $w(\alpha)$ is the value the statistic would need to exceed and depends on the distribution of the statistic and on how accurate of a test we want. Typically, α is taken to be 0.05. For example, if a statistic Z is standard normally distributed in the limit and we have significance level $\alpha = 0.05$, we would reject the null hypothesis if $Z > 1.645$. 1.645 is also known as the 95th percentile of the normal distribution since $\mathbb{P}(Z > 1.645) = 0.05 = \alpha$.
Back to our test statistics(the CLRT, CJ and LS); They are not standard normally distributed, but still normal. We can centralize any normal variable into standard normal however by subtracting the mean and dividing by the standard deviation. That would then be our rejection condition: For any statistic T with mean μ and variance σ^2:

$$\text{Reject } H_0 \text{ if } \frac{T-\mu}{\sigma} > 1.645$$

We proceed in the following manner:

1. Generate the sample covariance matrix S_n
2. Compute the three statistics for the CLRT, CJ and LS test
3. Check if the centralized statistics exceed $w(\alpha)$

This is a Bernoulli experiment. Depending on what hypothesis we condition, we have either success parameter $p = \mathbb{P}(\text{reject } H_0 \mid H_0 \text{ is true})$ or $p = \mathbb{P}(\text{reject } H_0 \mid H_1 \text{ is true})$. Recall that we had defined the following definition for power of a test:

$$\text{Power} = \mathbb{P}(\text{reject } H_0 \mid H_1 \text{ is true})$$

The more a powerful a test is, the quicker it rejects a false hypothesis. It is of our interest to compute and compare the empirical powers of the CLRT, CJ and LS test.
In our case we repeat above procedure 10000 times. By the weak law of large numbers we have that for our Bernoulli trial under H_1:

$$\frac{\text{amount of times we reject } H_0}{n} \xrightarrow{\mathbb{P}} \mathbb{P}(\text{reject } H_0 \mid H_1 \text{ is true}), n \to \infty$$

4.1 Empirical sizes

Before we go into the empirical powers we quickly take a look at the empirical sizes. If our statistics are close to the limiting distribution, then the empirical size of the test should be close the rejection level $\alpha = 0.05$ under H_0. Recall that the definition for size of a test was $\mathbb{P}$(reject $H_0 \mid H_0$ is true). We will simulate again 10000 times times a sample covariance matrix under H_0, using either standard normal data or $Gamma(4, 2) - 2$ variables.

(p, n)	CLRT	CJ	LS
(4,128)	0.0591	0.0585	0
(8,128)	0.0567	0.0556	0.0005
(16,128)	0.0518	0.0515	0.0123
(32,128)	0.0536	0.0535	0.0263
(64,128)	0.0533	0.0533	0.0409
(96,128)	0.0569	0.0508	0.0423
(112,128)	0.0536	0.0526	0.0459
(128,128)	0	0.0481	0.0413
(8,256)	0.0543	0.0538	0
(16,128)	0.0531	0.0523	0.0115
(32,128)	0.0551	0.0545	0.0284
(64,256)	0.0486	0.0528	0.0376
(96,256)	0.0478	0.0449	0.0377
(128,256)	0.0467	0.0486	0.0425
(192,256)	0.0503	0.0476	0.0422
(224,256)	0.054	0.0509	0.0470
(256,256)	0	0.0521	0.0488

(a) N(0,1) data

(p, n)	CLRT	CJ	LS
(4,128)	0.0741	0.0745	0
(8,128)	0.0731	0.0802	0
(16,128)	0.0636	0.0703	0.0461
(32,128)	0.0563	0.0654	0.0231
(64,128)	0.0491	0.0533	0.0323
(96,128)	0.0504	0.0588	0.0404
(112,128)	0.0487	0.0573	0.0416
(128,128)	0	0.0586	0.0436
(8,256)	0.0543	0.0538	0
(16,128)	0.0531	0.0523	0.0115
(32,128)	0.0551	0.0545	0.0284
(64,256)	0.486	0.0528	0.0376
(96,256)	0.0478	0.0449	0.0377
(128,256)	0.0476	0.0486	0.0425
(192,256)	0.0503	0.0472	0.422
(224,256)	0.0541	0.0509	0.470
(256,256)	0	0.0521	0.0488

(b) Gamma(4,2)-2 data

Table 4.1: Empirical sizes for the CLRT, CJ and LS test drawn from standard normal and Gamma(4,2)-2 variables

For most tests, we observe that the empirical sizes are close to the significance level, except for LS. One thing that immediately stands our in the table is that the sizes of the LS test are really small if p is small compared to n. This power increases as p grows larger. Therefore we can already conclude somewhat that the LS statistic relies more on the limiting aspect than the CJ test or CLRT test. As expected, the CLRT test has power 0 if $p = n$.

4.2 Empirical power comparisons

In this section we will try to find empirical powers for our three statistics. We will consider three different alternative hypothesis:

- H_1 : Equicorrelation
- H_1 : Auto-regressive
- H_1 : Fixed ratio of variables have variance $\neq 1$

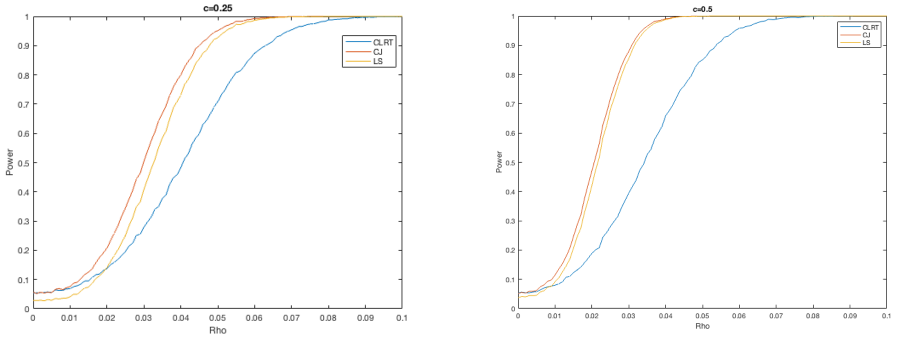
Equicorrelation

The first alternative we will investigate is an equicorrelation alternative. For this we are taking a linear combination of the identity matrix, and a matrix of all ones. This combination tests the alternative that the parameters of the underlying data have variance 1, but have covariance $Cov = \rho$ with other variables, so they are all correlated and thus dependent. For $\rho \in (0, 1)$, we define our alternative hypothesis as:

$$H_1 : \Sigma_{n,\rho} = (1-\rho)I + \rho\Delta^T\Delta, \Delta = (1,1,\dots,1),\text{(length p)}$$

$$\Sigma_{n,\rho} = (1-\rho)\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & \\ 0 & \dots & & 1 \end{bmatrix} + \rho \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & & \vdots \\ \vdots & & \ddots & \\ 1 & \dots & & 1 \end{bmatrix} = \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & & \vdots \\ \vdots & & \ddots & \\ \rho & \dots & & 1 \end{bmatrix}$$

We let ρ run from 0 to 1. This is also what we mean with increasing the *distance* of the alternative hypothesis. As ρ increases, the covariance matrix Σ_n becomes less and less like the identity matrix, our null hypothesis. For each ρ, we get a different alternative hypothesis, and thus we get different powers for each test. We now generate our sample covariance matrix $S_n = \frac{1}{n}YY^T$ with $Y = \Sigma_{n,\rho}^{1/2}X$. We take values for $p = 32, 64$ and $n = 128$, so $c = 1/4, 1/2$. For demonstrating purposes, we assume these parameters to be large enough such that the limiting distributions of the statistics are present.



(a) Empirical powers for $0 < \rho < 0.1, p = 32$ (b) Empirical powers for $0 < \rho < 0.1, p = 64$

In the figures we see how the power of the tests change when ρ increases. We see that already quite quickly for small ρ, around about 0.08 all tests have power 1. An explanation for that is that equicorrelation is quite a strong assumption. Even for small ρ, the test will quite quickly reject the null hypothesis. We also notice that the CLRT test behaves marginally worse than the CJ and the LS test.

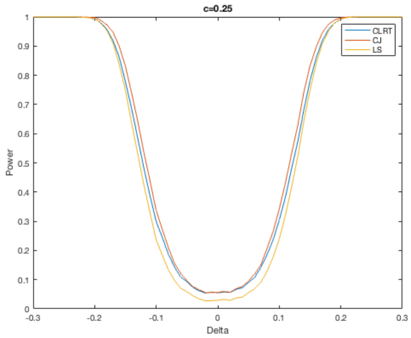
As expected, the CJ and LS test behave nearly the same, especially for $p = 64$. However, we notice a small discrepancy, which is even larger if $p = 32$. This is likely due to a small head start in power that the CJ test has over the LS test. In any case, both CJ and LS outperform the CLRT test.

Auto-regressive

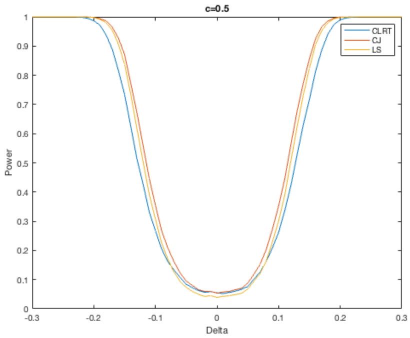
The second alternative we will compare is an auto-regressive relation. For any $\delta \in \mathbb{R}$, we define the element on the i-th row and jth column of the $p \times p$ true covariance matrix as $\delta^{|i-j|}$:

$$H_1 : \Sigma_{n,\delta} = \begin{bmatrix} 1 & \delta & \delta^2 & \dots & \delta^{p-1} \\ \delta & 1 & \delta & \dots & \delta^{p-2} \\ \vdots & \delta & \ddots & & \vdots \\ & & & \ddots & \delta \\ \delta^{p-1} & \delta^{p-2} & \dots & \delta & 1 \end{bmatrix}$$

We could pick any $\delta \in \mathbb{R}$, but we restrict ourselves to $\delta \in (-1, 1)$, since this corresponds with a stationary auto-regressive model. And we will see shortly, rejection of the null hypothesis happens for small δ very quickly. We again generate the sample covariance matrix $S_n = \frac{1}{n}YY^T$ with $Y = \Sigma_{n,\delta}X$. We take values for $p = 32, 64$ and $n = 128$.



(a) Empirical powers for $-0.2 < \delta < 0.2$ p = 32



(b) Empirical powers for $-0.2 < \delta < 0.2$ p = 64

Under this hypothesis, the three tests seem to behave quite comparable. For $p = 64$, the CLRT performs just a tiny bit worse from $\delta = 0.1$ and onwards. However, when $p = 32$ the CLRT behaves nearly as good as the CJ test while the LS test performs worse. This agrees with an observation from before regarding the sizes of the test. When p is small, and the distance to the null hypothesis is not as big yet, the LS test performs very poorly compared to the CLRT and the CJ test. It starts at a lower point on the power curve but picks up as δ departs from 0.

This test needs a larger parameter δ compared to the ρ test to pick up power since for small δ, powers of δ will be exponentially smaller and thus for small δ our alternative is closer to the null hypothesis than if we had the same parameter value for ρ. That is why the tests are at power 1 from $\delta = 0.2$, whereas for ρ it had power 1 from $\rho = 0.08$ and onwards.

We notice that the power curve is symmetric about 0. Explanation for that is our definition for $Y = \Sigma_{n,\delta}^{1/2} X$. The true covariance matrix $\Sigma_{n,\delta}$ is a covariance matrix and thus positive definite, and by theorem 1.1.4 $\Sigma_{n,\delta}$ has positive eigenvalues, and the square root matrix exists and is real. It may have negative elements, but that does not matter since the elements of X are standard normal, and thus we get positive elements as often as we get negative elements in Y, regardless of the parity of δ.

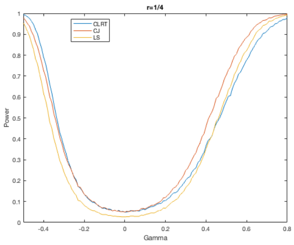
Fixed ratio with variance other than 1

The third alternative we consider is an alternative where a fixed ratio r of the variables have a variance not equal to 1, but equal to $1 + \gamma$. In particular, for any $\gamma \in \mathbb{R}$ we define our alternative hypothesis

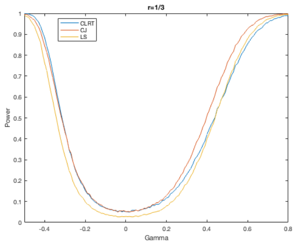
$$H_1 : \Sigma_{n,\gamma,r} = \begin{bmatrix} 1 & 0 & & \dots & 0 \\ 0 & \ddots & & & \vdots \\ & & 1+\gamma & & \\ \vdots & & & \ddots & \\ 0 & \dots & & & 1+\gamma \end{bmatrix}$$

For example if we pick a ratio of $r = 1/2$, then half of the p variables will have variance $1 + \gamma$, in our alternative hypothesis. If it happens that $r \times p$ is not a whole number, we will round it down. For example, if $r = 1/3$ and $p = 64$, then we consider 21 non-unit variance random variables, at the end of the diagonal.

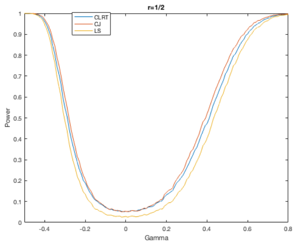
To differentiate between the effects of varying the ratio versus varying the dimension, we consider the ratios $\frac{1}{2}, \frac{1}{3}$ and $\frac{1}{4}$, for both $p = 32$ and $p = 64$.



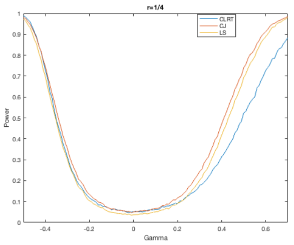
(a) $p = 32$, ratio $= 1/4$



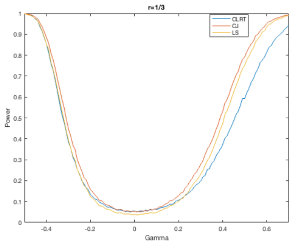
(b) $p = 32$, ratio $= 1/3$



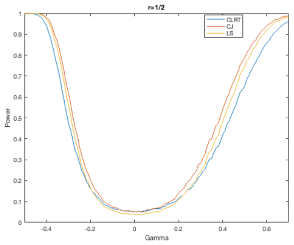
(c) $p = 32$, ratio $= 1/2$



(a) $p = 64$, ratio $= 1/4$



(b) $p = 64$, ratio $= 1/3$



(c) $p = 64$, ratio $= 1/2$

For $p = 32$, the tests behave comparable just like the auto-regressive case. Their powers remain relatively close to each other, especially when $r = \frac{1}{2}$. However it seems that the CJ test performs the best of the three across the board.

When $p = 64$, again we see that the CJ and LS test outperform the CLRT again, this time by a bit larger margin. However as the ratio increases the CLRT starts performing a little bit better, as expected since the distance from the null hypothesis is larger for a bigger ratio r. Again, the CJ test and LS test perform very similar, but CJ is slightly better still.

4.3 ROC curves

Another way of comparing the quality of tests are receiver operating characteristic curves, or ROC curves for short. These are curves that compare how the true positive rate changes as we vary our desired false positive rate. ROC curves are a type of power plot, but instead of varying the underlying distance variable, we vary the significance level for when we reject. The true positive rate is defined as the probability that a test correctly rejects a false hypothesis: $\mathbb{P}(\text{reject } H_0 \mid H_1 \text{ is true})$. We recognize here the power of the test. These curves allow us to assess the quality of the tests when their power are close to each other.

We proceed in the following manner. We pick a value of either ρ, δ and γ under an alternative hypothesis where the powers are close. However, we also want to pick a value that is as far away from 0 as possible, otherwise we are too close to the H_0 hypothesis and then our test isn't worth much. So we try to find the empirical value such that the powers are as close as possible, but the parameter should be as large as possible. We let our false positive rate α run from 0 to 1, instead of fixing it at 0.05.

Now our rejection condition changes; First we rejected if the centralized test statistics exceeded 1.645, which corresponded with a significance level of $\alpha = 0.05$. We can compute now for different α the value $w(\alpha)$, the value which a normal statistic has to exceed to be in the $\alpha\times$ 100th percentile via the relation:

$$
\begin{aligned}
1-\alpha &= \frac{1}{2\sqrt{\pi}}\int_0^{w(\alpha)} e^{\frac{-t^2}{2}}\,dt \\
\iff 1-\alpha &= \frac{1}{\sqrt{\pi}}\int_0^{w(\alpha/\sqrt{2})} e^{-t^2}\,dt \\
\iff 1-\alpha &= \frac{1}{2}+\frac{1}{2}erf(w(\alpha)/\sqrt{2}) \\
\iff w(\alpha) &= \sqrt{2}\cdot erfinv(1-2\alpha)
\end{aligned}
$$

Here we used the error function $erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$, and it's inverse which can be computed numerically.

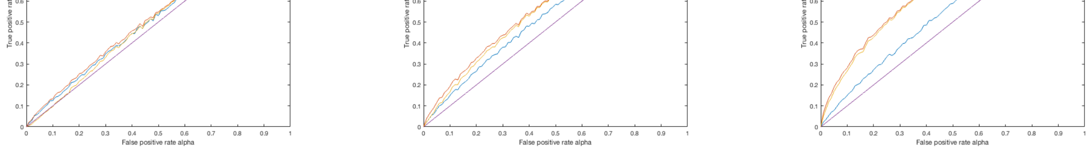
Recall our initial rejection condition: $\frac{T-\mu}{\sigma} > w(\alpha)$.
If we insert $\alpha = 0.05$ into this expression, one would find $w(\alpha) = 1.645$. If we insert $\alpha = 0$, we would find $w(0) = \infty$. We interpret that the following way: If $\alpha = 0$, we think that no test outcome is significant at all. Since our test statistic can never be greater than infinity, we never reject the null hypothesis no matter what hypothesis is true and we get power 0 for all tests. As we allow α to steadily grow, depending on the test we will get different but increasing powers. If α reaches 1, that corresponds with a belief that every test outcome is significant, so you always reject. We again see that back in the $w(\alpha)$ value, which is $w(1) = -\infty$. No matter what simulation, every statistic exceeds this and thus always leads to a rejection, thus power 1.

We compute $w(\alpha)$ and use this value to check if our centralized statistic exceeds this value. Again, we run it 10000 times and compute the empirical probability, just like for the power plots. If a test gains more power than the other at the same significance level, than that test could be considered as a better test.

One thing we also include as a comparison is a straight line, which represents a test that we will refer to as ”standard”. This test is a test that given a significance level α as it's false positive rate, has also true positive rate α. The better a test is, the further away from this straight line it should be.

ROC curves for equicorrelation

Looking in figure 4.1b, we see that the three tests are (as expected) most comparable to each other when ρ is small. A value we could use would be $\rho = 0.01$. That gives us the following ROC curves, where in the first case we have $p = 32$, so $c = 1/4$, in the second case $p = 64$ and in the other we have $p = 112$ so $c = 7/8$. This is to check whether or not c is of very relevant importance. Our only test that directly depends on the value of c was the CLRT test. However, our other tests ”rely” on the value of c in the sense that the higher we have c, for fixed n, the value of p is relatively larger so the limiting behaviour should be more present and thus lead to higher power.



(a) $p = 32$

(b) $p = 64$

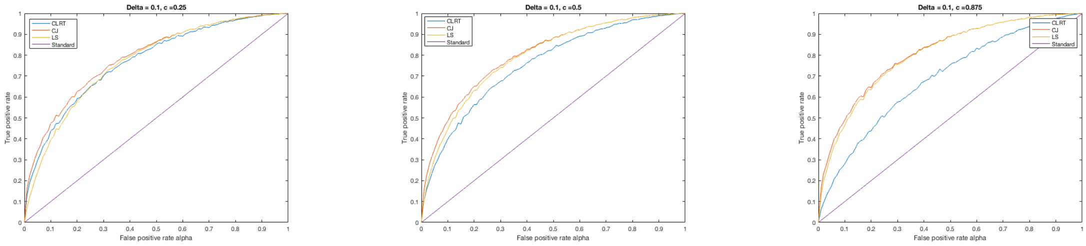
(c) $p = 112$

If $p = 32$, then our tests are not powerful, for this $\rho = 0.01$. It seems that for this case, the high-dimensional aspect is not able to make up for the still relative small distance to H_0.
But we see that as our p increases, the CJ and LS test outperform the CLRT, by a significant margin. As a matter of fact, CLRT only increases a small bit, while CJ and LS performance

increases well as p increases. As said earlier, this could be due to the limiting behaviour of CJ test and LS test coming out stronger as p increases.

ROC curves, auto-regressive

We pick, similarly to ρ, a value of $\delta \neq 0$ where the tests are comparable. A good value we could try would be $\delta = 0.1$. We again repeat the process for $p = 32, 64, 112$.



(a) $p = 32$

(b) $p = 64$

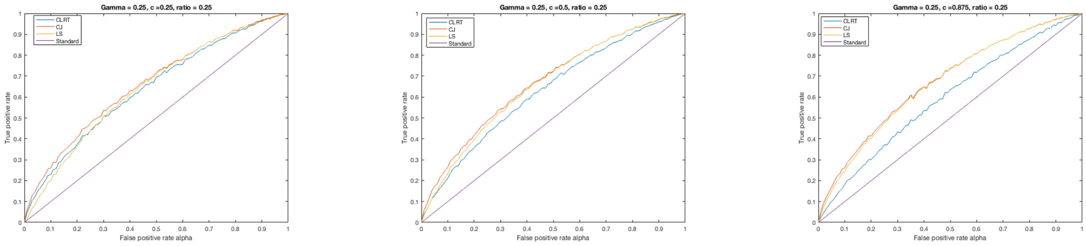
(c) $p = 112$

Here we see something different happening from the equicorrelation case. Instead of increasing in power as p increases, all tests are already powerful quite quickly.

However in this case, CLRT now performs increasingly worse, while CJ and LS stay at their power level. The CLRT test performs better for equicorrelation tests where p becomes larger, and falls of in power in the auto-regressive case.

ROC curves, ratio case

The γ case is a bit more interesting, since we have a factor extra, namely the ratio that dictates how large a portion of the diagonal has $1 + \gamma$ variance. Looking at the graphs for power, a value of γ where the powers are close could be $\gamma = 0.25$. We fix the ratio at $\frac{1}{4}$ and $\frac{1}{2}$



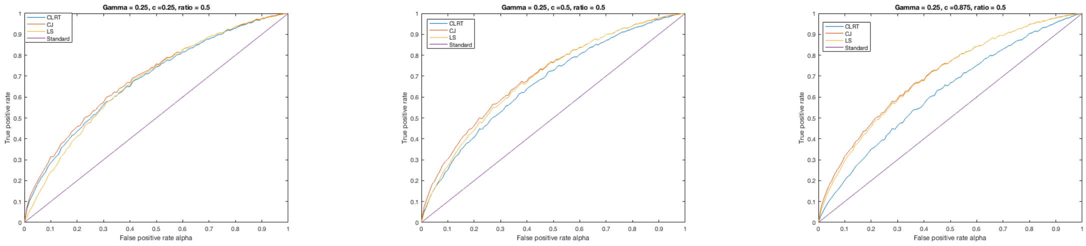
(a) Ratio = 1/4 , $p = 32$

(b) Ratio = 1/4 , $p = 64$

(c) Ratio = 1/4 , $p = 112$

Here something similar to the auto-regressive case is happening. For $p = 32$, the three tests behave similarly, but the CLRT performs worse as p increases.

Now for $r = \frac{1}{2}$



(a) Ratio = 1/2 , $p = 32$

(b) Ratio = 1/2 , $p = 64$

(c) Ratio = 1/2 , $p = 112$

Compared to $r = \frac{1}{4}$, all the three test are a bit more curved outwards. This is logical, since bigger ratio means larger distance to the null hypothesis and thus quicker rejection. We also again see the CLRT test decreasing in power.

4.4 Conclusions

In most curves, we see a trend. The CLRT statistic performs worse, or sometimes equal to the CJ and LS test. This is something I had not initially expected, since the likelihood ratio test is a good test in finite dimensional statistics.

Something else we see is that while CJ and LS test perform similarly, the CJ test just slightly outperforms the LS test, especially when our alternative is close the null hypothesis. For example, notice the discrepancy between the CJ and LS statistic in figure 4.1a. It seems like the CJ test has a small head start in power compared to LS. As ρ increases, both these tests increase in power, but the discrepancy never truly disappears until the power is 1.
A similar observation is made in figure 4.2a. Our LS test starts lower, and is never truly able to catch up to CLRT and CJ.

I think this phenomenon has the same explanation as why the sizes of the LS are so small. The LS test needs the high-dimensional aspect more than CLRT and CJ do. One could argue that for power curves, $p = 32$ is too small for high-dimensional behaviour to really come forward, and use that as an explanation for why the LS test is set back a little compared to the CJ test. I'm inclined to agree with that, since all of these derivations were done under the assumption that $p \to \infty$, and we did see LS test have higher sizes if p was about the same as n, in table 4.1.

However, I would like to point a small rule of thumb that is often used in practice: The regular one dimensional Central Limit Theorem is in practice applicable for sample sizes of n = 30, if your data is roughly bell-shaped. In our case we used standard normal data so this satisfies the condition. I would expect that $p = 32$ is large enough to at least apply our Central Limit Theorem 1.5.1 for linear spectral statistics.

It seems like LS is never really able to catch up to the CJ because of the "delay", the difference in power, it starts with. A positive note is that under our alternative hypothesis, it still mostly outperforms the CLRT. Based on my findings, I would recommend to use the CJ test over the LS and CLRT test at all levels, either finite or high-dimensional statistics. That is because LS only works good in high-dimensional case, but even then it just has the same distribution as the CJ test, and so produces the same results. The CLRT test performs decent in for low dimension, but falls short as p increases.

Chapter 5

Non-linear shrinkage estimators

We now consider a new type of estimator; the non-linear shrinkage estimator by Ledoit and Wolf [12]. The aim of this section is to find out if we can proceed in a similar way as we did for linear shrinkage; that is find parameters that rely on the sample eigenvalues in some way and find a limiting distribution. If we are able to do that, we can construct a new statistical test that might be able to outperform the linear shrinkage, or the CJ-test. Before we are potentially able to find out a test we first need to lay down the framework for non-linear shrinkage.

5.1 Setting

We again have the following setting:
We are interested in the spectral properties of the sample covariance matrix

$$S_n = \frac{1}{n} Y Y^T$$

where $Y = \Sigma_n^{1/2} X_n$, satisfying the following assumptions:

- (A1) Our $p \times n$ noise matrix X_n consists of independent identically distributed variables $x_{i,j}$ with mean 0 and variance 1, and the 12-th moment $\mathbb{E}x_{(i,j)}^{12}$ exists.
- (A2) Our true population covariance Σ_n is a $p \times p$ random matrix, independent of X_n
- (A3) $\frac{p}{n} \to c \in (0, 1)$
- (A4) $\tau_1 \dots \tau_p$ are the eigenvalues of Σ_n, and the empirical spectral distribution $H_n(\tau)$ converges to some non-random limit law $H(\tau)$

Before presenting the results we briefly present some definitions and known results that we will need for our non-linear shrinkage case.

Definition 5.1.1. [2] Let F be a finite measure on the real line. The *Stieltjes transform* of the measure F with $z \in \mathbb{C}^+ : \{z \in \mathbb{C} : Im(z) > 0\}$ is defined as

$$m_F(z) = \int \frac{1}{x - z} dF(x)$$

Let $(\lambda_1 \dots, \lambda_p)$ denote the eigenvalues of S_n, in decreasing order. If the measure is the empirical spectral distribution F^{S_n} of the eigenvalues of S_n, one gets for it's Stieltjes transform

$$m_{F_n}(z) = \frac{1}{p} \sum_{i=1}^{p} \frac{1}{\lambda_i - z} = \frac{1}{p} tr((S_n - zI)^{-1})$$

The first major result is by Marcenko-Pastur (1967) [13]. In the next theorem we recall their result and present it the most recent version as given in [14].

Theorem 5.1.1. *[13] Let $m_{F_n}(z) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{\lambda_i - z}$ be the Stieltjes transform of the ESD of S_n. Then, under assumptions (A1)-(A4) we have for $\forall z \in \mathbb{C} : Im(z) > 0$, that $\lim_{n\to\infty} m_{F_n}(z) = m_F(z)$ almost surely, where $m_F(z)$ satisfies:*

$$\forall z \in \mathbb{C}^+ : m_F(z) = \int_{-\infty}^{\infty} \left\{\tau[1 - c - czm_F(z)] - z\right\}^{-1} dH(\tau)$$

As always, the empirical distribution function $F_p(x) = \frac{1}{p}\sum_{i=1}^{p} 1_{[\lambda_i < x]}$ of the sample covariance matrix converges to the nonrandom limit F, the cumulative distribution function of the Marcenko-Pastur law.

In addition, it is proven in [15] that the following limit exists:

$$\lambda_i \in \mathbb{R} \setminus 0 : \lim_{z \in \mathbb{C}^+ \to \lambda_i} m_F(z) := \breve{m}_F(\lambda_i)$$

Here it comes into play why we required that $c \in (0, 1)$. If $c > 1$, the sample covariance matrix has eigenvalues equal to zero, with proportion $1 - \frac{1}{c}$. We can then not always be sure if this limit exists in this current form, so we stick to $c < 1$.

The general form for the non-linear shrinkage estimator we will consider is:

$$\Sigma_{NLS} = U_n D_n U_n^{-1}$$

Here $D_n = Diag(d_1, \ldots, d_p)$ is a diagonal matrix, and U_n is the matrix consisting of the sample eigenvectors of S_n. We again want to minimize the following expression:

$$min_{D_n} ||U_n D_n U_n^{-1} - \Sigma_n||_F^2, D_n \text{ diagonal}$$

It can be shown [11] that the optimal solution is

$$\tilde{D_n} = diag(\tilde{d_1}, \ldots, \tilde{d_p}), \forall i \in \{1, \ldots, p\} : \tilde{d_i} = u_i^T \Sigma_n u_i$$

where u_i denotes the eigenvector corresponding to the i-th eigenvalue. We cannot explicitly calculate these $\tilde{d_i}$ since they depend on the non-observable Σ_n. However, the following theorem, Theorem 4 from [11] gives us the asymptotic quantity that will estimate the non-linear shrinkage intensity.

Before we can state the theorem we need to introduce the following object:

$$\forall x \in \mathbb{R}, \Delta_p(x) = \frac{1}{p}\sum_{i=1}^{p} \tilde{d_i} \cdot 1_{\{\lambda_i < x\}}$$

This is a type of cumulative distribution with jumps of height $\tilde{d_i}$ at the eigenvalues. This distribution function will settle down to some non-random cumulative function Δ:

Theorem 5.1.2. *Assume conditions (A1)-(A4) hold, and let $\Delta_p(x)$ be defined as above. Then there exists a non-random function Δ defined on $\mathbb{R}$ such that $\Delta_p(x)$ converges almost surely to Δ for all $x \in \mathbb{R} \setminus \{0\}$. If furthermore $c < 1$, then Δ can be written in the form $\forall x \in \mathbb{R}, \Delta(x) = \int_{-\infty}^{x} \delta(\lambda) dF(\lambda)$, where*

$$\delta(\lambda) = \frac{\lambda}{|1 - c - c\lambda\breve{m}_F(\lambda)|^2}, \lambda > 0$$

So the asymptotic quantity that corresponds with $\tilde{d_i}$ is $\delta(\lambda_i)$, which is the sample eigenvalue divided by a correction factor $|1 - c - c\lambda\breve{m}_F(\lambda_i)|^2$.

5.2 Non-linear shrinkage intensities under H_0

Now that we know what our optimal non-linear shrinkage estimator looks like, we would like to find out if we can produce some tests. However, this is not possible as a result of the following theorem:

Theorem 5.2.1. *Let $\delta(\lambda_i)$ be the transformed sample eigenvalue defined as in Theorem 5.1.2, with $\lambda_i \in [a,b] = [(1 \mp \sqrt{c})^2]$, the support of the Marcenko-Pastur distribution. Then under $H_0 : \Sigma_n = I$, $\delta(\lambda_i) = 1$.*

Proof. Recall that $\forall i = 1 \dots p : \delta(\lambda_i) = \frac{\lambda}{|1-c-c\lambda \check{m}_F(\lambda_i)|^2}$. Computing $\delta(\lambda_i)$ is thus directly equivalent with computing $\check{m}_F(\lambda_i)$. Recall it's definition:

$$\lambda_i \in \mathbb{R} \setminus 0 : \check{m}_F(\lambda_i) \coloneqq \lim_{z \in \mathbb{C}^+ \to \lambda_i} m_F(z)$$

Recall that $m_F(z)$ satisfied

$$m_F(z) = \int_{-\infty}^{\infty} \left\{ \tau[1-c-czm_F(z)] - z \right\}^{-1} dH(\tau)$$

We first need to find a useable expression for $m_F(z)$ before we can apply the limit. We use that under H_0, the true covariance matrix is $\Sigma_n = I$. This matrix has as eigenvalue $\lambda = 1$, with multiplicity p. That means that the underlying $H(\tau)$, the distribution function of the true eigenvalues, jumps to 1 at $\tau = 1$. This means that the integral with respect to $H(\tau)$ under H_0 is:

$$m_F(z) = \int_{-\infty}^{\infty} \left\{ \tau[1-c-czm_F(z)] - z \right\}^{-1} dH(\tau) = \frac{1}{1-c-czm_F(z)-z}$$

Rewriting terms gives us a quadratic equation in $m_F(z)$:

$$czm_F(z)^2 + m_F(z)(c+z-1) + 1 = 0$$

Applying the quadratic formula gives us as solutions:

$$m_F(z) = \frac{(1-c) - z \pm \sqrt{(c^2+z^2+1-2c-2z+2cz) - 4c}}{2cz}$$

We restrict us to the case where we take the "plus" solution, but as we will see shortly this does not matter. The term withing the square root can be simplified to $\sqrt{(b-z)(a-z)}$, with $b = [1+\sqrt{c}]^2$ and $a = [1-\sqrt{c}]^2$ the boundaries of the support of the eigenvalues. We have to take the limit as $z \to \lambda_i$ of the above expression, which has reduced to

$$m_F(z) = \frac{(1-c) - z + \sqrt{((b-z)(a-z)}}{2cz}$$

Since we are working with complex numbers we have to define a branch cut for our square root. We take the principal branch with $Arg(z) \in (-\pi, \pi]$ and let

$$\sqrt{z} = \exp(\log(z^{\frac{1}{2}}) = \exp(\frac{1}{2}(\ln|z| + i \cdot Arg(z))) = \exp(\frac{1}{2}\ln|z|)\exp(\frac{i}{2}Arg(z))$$

Here i denotes the imaginary unit that satisfies $i^2 = -1$.

Since our eigenvalues are strictly positive (since $c < 1$) we won't have trouble with the fact that the square root is undefined in $z = 0$. This square root is then continuous and we can happily take our limit $z \to \lambda_i$. This gives:

$$\check{m}_F(\lambda_i) = \lim_{z \to \lambda_i} m_F(z) = \frac{(1-c) - \lambda_i + \sqrt{(b-\lambda_i)(a-\lambda_i)}}{2c\lambda_i}$$

Inserting this into our $\delta(\lambda_i)$ gives

$$\begin{aligned}
\delta(\lambda_i) &= \frac{\lambda_i}{|1-c-c\lambda_i \check{m}_F(\lambda_i)|^2} \\
&= \frac{\lambda_i}{|1-c-c\lambda_i(\frac{(1-c)-\lambda_i+\sqrt{(b-\lambda_i)(a-\lambda_i)}}{2c\lambda_i})|^2} \\
&= \frac{\lambda_i}{|(1-c) - \frac{1}{2}(1-c) + \frac{\lambda_i}{2} - \frac{1}{2}\sqrt{(b-\lambda_i)(a-\lambda_i)}|^2} \\
&= \frac{\lambda_i}{|\frac{1}{2}(1-c) + \frac{\lambda_i}{2} - \frac{1}{2}\sqrt{(b-\lambda_i)(a-\lambda_i)}|^2}
\end{aligned}$$

Now we notice that since $a < \lambda_i < b$, the expression in the square root is negative and real. We can however write this as

$$\sqrt{(b-\lambda_i)(a-\lambda_i)} = \sqrt{(-1)(b-\lambda_i)(\lambda_i - a)} = i\sqrt{(b-\lambda_i)(\lambda_i - a)}$$

Here we used the principal branch of our square root, that maps strict negative reals onto the positive imaginary axis, without 0. That gives for our $\delta(\lambda_i)$:

$$\delta(\lambda_i) = \frac{\lambda_i}{|\frac{1}{2}(1-c) + \frac{\lambda_i}{2} - \frac{i}{2}\sqrt{(b-\lambda_i)(\lambda_i - a)}|^2}$$

We now work out the denominator. For any complex z it holds that $|z|^2 = \Im(z)^2 + \Re(z)^2$. For the denominator we have that it's real part is $\frac{1}{2}(1-c+\lambda_i)$, and imaginary part $-\frac{1}{2}\sqrt{(b-\lambda_i)(\lambda_i - a)}$. Here we notice from earlier that it does not matter if we took the "plus" or "minus" solution for $m_F(z)$, because we take the imaginary part squared and the "plus" or "minus" only determined the sign of the imaginary part. We get for the real part and imaginary part of the denominator, writing ω for short:

$$\begin{aligned}
\Re(\omega)^2 &= (\frac{1}{2}(1-c+\lambda_i))^2 \\
&= \frac{1}{4}(1-c+\lambda_i)^2 \\
&= \frac{1}{4}(1+c^2+\lambda_i^2 - 2c + 2\lambda_i - 2c\lambda_i) \\
\Im(\omega)^2 &= (-\frac{1}{2}\sqrt{(b-\lambda_i)(\lambda_i - a)}^2 \\
&= \frac{1}{4}((b-\lambda_i)(\lambda_i - a)
\end{aligned}$$

We separately write out $(b-\lambda_i)(\lambda_i - a)$

$$\begin{aligned}
(b-\lambda_i)(\lambda_i - a) &= ([1+\sqrt{c}]^2 - \lambda_i)(\lambda_i - [1-\sqrt{c}]^2) \\
&= (1+2\sqrt{c}+c-\lambda_i)(\lambda_i - 1 + 2\sqrt{c} - c) \\
&= \lambda_i - 1 + 2\sqrt{c} - c + 2\lambda_i\sqrt{c} - 2\sqrt{c} + 4c - 2c\sqrt{c} \\
&+ c\lambda_i - c + 2c\sqrt{c} - c^2 - \lambda_i^2 + \lambda_i - 2\lambda_i\sqrt{c} + c\lambda_i \\
&= 2c + 2c\lambda_i + 2\lambda_i - c^2 - 1 - \lambda_i^2
\end{aligned}$$

Taking this all together gives us for the denominator of $\delta(\lambda_i)$

$$\begin{aligned}
\Re(\omega)^2 + \Im(\omega)^2 &= \frac{1}{4}(1+c^2+\lambda_i^2 - 2c + 2\lambda_i - 2c\lambda_i) \\
&+ \frac{1}{4}(2c + 2c\lambda_i + 2\lambda_i - c^2 - 1 - \lambda_i^2) \\
&= \frac{1}{4}(4\lambda_i) = \lambda_i
\end{aligned}$$

Now recalling the definition of $\delta(\lambda_i)$, we get that both its enumerator and denominator are λ_i, and thus we get that for all eigenvalues $\lambda_1 \ldots \lambda_p$, it's transformed optimal non-linear shrinkage intensity $\delta(\lambda_i) = 1$. □

This mere fact makes it not possible to construct tests in the same manner as we did for linear shrinkage. For linear shrinkage, the fact that made constructing tests possible was that the optimal intensity, α^*, depended on linear spectral statistics. These statistics, like the trace $\sum_{i=1}^{p} \lambda_i$ or squared Frobenius norm $\sum_{i=1}^{p} \lambda_i^2$ applied one analytic function to all the the eigenvalues, using $f(x) = x$ and $f(x) = x^2$ respectively. This allowed us to use the central limit theorem 1.5.1, since that theorem stated that functionals were normally distributed in the high-dimensional case.

In the non-linear shrinkage case, we apply a function $\delta(\lambda)$ to each eigenvalue. However, it turns out that this function $\delta(\lambda_i)$ maps each eigenvalue to 1 under H_0. In one sense this is good news. Our true covariance matrix *is* the identity matrix, so it has eigenvalues 1, and the non-linear shrinkage $\delta(\lambda_i)$ turns the eigenvalues perfectly into what the true eigenvalues actually are. It makes it however nonrandom.

The fact that it turns the eigenvalues into non-random quantities is what makes testing as we did for linear shrinkage not possible. For those tests we were able to find a distribution under H_0 and from that distribution we could define a rejection condition for significance levels. For non-linear shrinkage this would not make sense since we would already know prior if our statistics would exceed a certain value. For example, lets say we would define a statistic $T = \frac{1}{p}\sum_{i=1}^{p}\delta(\lambda_i)$. Because $\delta(\lambda_i) = 1$ for all i, we would get that $T = \frac{1}{p} \cdot p \cdot 1 = 1$. This is a degenerate statistic, and tests using this would always result in power 0 or 1 and size 0 or 1.

Chapter 6

Conclusions

In this concluding section I would like to review my work and draw some conclusions from it. I will also discuss some points where I think I could have done better, and what further research one could do.

6.1 Conclusions

In this paper we have derived a new estimator based on the linear shrinkage estimator by Bodnar (2014) [1]. We found that when multiplied by it's dimension, the optimal estimator $\hat{\alpha}^*$ converges to a normal distribution, and that distribution is the same as the corrected distribution for the John's test, which itself was constructed as an optimal test under rotation invariance. After the derivation we have done simulation study and found out that the LS test was inferior to CJ test, by virtue of it being equally distributed in high-dimensional statistics and thus performing the same but worse in finite dimensional statistics.

We also tried to find a possible test in a similar way based on the non-linear shrinkage estimator by Ledoit (2011) [11]. However, it was not possible to directly construct a test due to the fact that under the null hypothesis the transformed eigenvalues were 1, and thus non random.

6.2 Discussion

The two tests I compared my LS test with, the CLRT and CJ test, are both initially constructed from a finite dimensional view. That is, these tests were originally meant for finite dimensional statistics, but were transformed by Wang and Yao in [3] to work in high-dimensional statistics as well. While both these tests are very powerful in finite dimensional statistics, these tests might not be the most powerful in high-dimensional case.

Some other test I could have used was from Onatski et. al (2013)[17]. Here they proposed another sphericity test that has high power in the high-dimensional case. It could have been interesting to see how our LS test compared to this test.

6.3 Recommendations

In this section I would like to briefly discuss what some things are that could be explored as a follow up

6.3.1 Prior belief Σ_0

As stated in section 3.1.1, our optimal estimator $\hat{\alpha}^*$ depends on our choice of prior belief Σ_0. Recall it's definition:

$$\hat{\alpha}^* = 1 - \frac{\frac{1}{n}||S_n||_{tr}^2||\Sigma_0||_F^2}{||S_n||_F^2||\Sigma_0||_F^2 - (tr(S_n\Sigma_0))^2}$$

We could of course try many more prior beliefs. However, this might come with some complications; The squared Frobenius can just be computed directly, but we might not be able to find an explicit distribution for $tr(S_n\Sigma_0)$. Dependent on how you define your Σ_0, this might not become linear spectral statistics and thus not fit for the Central Limit Theorem.

In Bodnar (2014)[1], we can find a handy theorem, theorem 3.2, that could help us out:

$$\frac{1}{p}|tr(S_nM) - tr(\Sigma_nM)| \xrightarrow{a.s} 0, (p,n) \to \infty, \frac{p}{n} \to c$$

Here M is any symmetric positive definite matrix. So as long as we choose a symmetric positive definite matrix Σ_0, we will be able to consistently estimate $tr(tr(\Sigma_n\Sigma_0)$ with $tr(S_n\Sigma_0)$. Dependent on what Σ_0 you choose, the quantity $tr(S_n\Sigma_0)$ might become linear spectral statistics or not. In either case we are able to consistently estimate it. As a result the optimal estimator $\hat{\alpha}^*$ multiplied by p may or may not be normal now.

6.3.2 Case when dimension is larger than sample size

In (nearly) all derivations we have done here, we assumed that $c \in (0,1)$, so our dimension is smaller than the sample size. The reason we did this was to avoid null eigenvalues. We already saw that the CLRT was not defined for $c \geq 1$, since it depended on $\log(1-c)$. However, a good part of the theory is still valid for $c > 1$. For example, the Central Limit Theorem 1.5.1 required only $c > 0$. A reason why this could be an interesting case is that in practice one works with a limited sample size, but you have an enormous amount of data per unit. For example, if you work in genetics you can measure the gene expression of a person, but this could incorporate a huge amount of genes. If you want to find out what the co-expressions are for certain genes you would need to construct a sample covariance matrix where $p > n$, or even $p >> n$, dependent on how many genes you need. Right now a commonly used technique is Principal Component Analysis. This method however throws away data. High-dimensional statistics could provide an answer, and researching the case $p > n \iff c > 1$ could be very interesting.

6.3.3 Non-linear shrinkage tests

In my research I was unable to construct a test for non-linear shrinkage. However, it is possible to construct a test in different ways. Li et al (2018) [16] proposed some tests based on non-linear shrinkage, like the Bartlett–Nanda–Pillai trace (BNP) test.

One could also proceed in a different way for non-linear shrinkage. Instead of taking the limit $z \to \lambda_i$ in $\lim_{z\to\lambda_i} m_F(z)$, one could try to take the limit as $z \to \lambda_i + \epsilon$, for some small $\epsilon > 0$, and then see what happens for the transformed eigenvalues. If we however simply add ϵ to the denominator in $\delta(\lambda) = \frac{\lambda}{|1-c-c\lambda\breve{m}_F(\lambda)|^2}$, one ends up with $\delta(\lambda) = \frac{\lambda}{\lambda+\epsilon}$. This is linear spectral statistics, and this we know what happens. You have that $\sum_{i=1}^{p} \frac{\lambda_i}{\lambda_i+\epsilon}$ is normal in the limit minus some correction factor. This could be interesting for future research.

Appendix A

Derivations of limiting distributions

A.1 Linear spectral statistics distribution derivation

In this section we will derive the limiting distributions for:

$$||S||_F^2 = \sum_{i=1}^{p} \lambda_i^2, Tr(S) = \sum_{i=1}^{p} \lambda_i$$

Throughout the whole section, $h = \sqrt{c}$.
A handy remark that will be useful in computing the contour integrals will be the following:

Remark. *Let $z \in \mathbb{C}$ run around the unit circle counterclockwise once, with complex conjugate $\bar{z} = 1/z$. Then*

$$\begin{aligned}
|1 + hz|^2 &= (1 + hz)(\overline{1 + hz}) \\
&= (1 + hz)(1 + \frac{h}{z}) \\
&= 1 + \frac{h}{z} + hz + h^2 \\
&= (z + h)(\frac{1}{z} + h) = \frac{(z + h)(1 + hz)}{z}
\end{aligned}$$

A.1.1 Limiting distribution of $||S_n||_F^2$

The squared Frobenius norm is a linear spectral statistic with $f = x^2$. Namely:

$$||S||_F^2 = tr(S^2) = \sum_{i=1}^{p} \lambda_i^2$$

Using the Central Limit Theorem 1.5.1, we get that under H_0:
$X_n(x^2) = p\{\frac{1}{p} \sum_{i=1}^{p} \lambda_i^2 - F_c(x^2)\}$ converges to a normal variable.
But $p(\frac{1}{p} \sum_{i=1}^{p} \lambda_i^2) = \sum_{i=1}^{p} \lambda_i^2$, which was our squared Frobenius norm. If we can find the limiting distribution of this expression, we are able to make statements about the Frobenius norm of our sample covariance matrix.

Expectation For the expectation we get if we apply the Central Limit Theorem 1.5.1 and Proposition 1.5.1:

$$\begin{aligned}
\mathbb{E}(X_{x^2}) &= (\kappa - 1)I_1(x^2) + \beta I_2(x^2) \\
&= (\kappa - 1)\lim_{r \downarrow 1} I_1(x^2, r) + \beta I_2(x^2) \\
&= (\kappa - 1)\lim_{r \downarrow 1} \frac{1}{2\pi i} \oint_{|z|=1} |1 + hz|^4 [\frac{z}{z^2 - r^{-2}} - \frac{1}{z}]dz + \beta \frac{1}{2\pi i} \oint_{|z|=1} |1 + hz|^4 \frac{1}{z^3} dz
\end{aligned}$$

We first compute the first contour integral, after $\kappa - 1$. From our remark we find that $|1+hz|^4 = (|1+hz|^2)^2$ is equal to $\frac{(z+h)^2(1+hz)^2}{z^2}$ This turns our contour integral into

$$\frac{1}{2\pi i}\oint_{|z|=1} \frac{(z+h)^2(1+hz)^2}{z^2}\Big[\frac{z}{z^2-r^{-2}} - \frac{1}{z}\Big]dz \tag{A.1}$$

$$=\frac{1}{2\pi i}\Big(\oint_{|z|=1} \frac{(z+h)^2(1+hz)^2}{z^2}\frac{z}{(z-\frac{1}{r})(z+\frac{1}{r})}dz - \oint_{|z|=1} \frac{(z+h)^2(1+hz)^2}{z^2}\frac{1}{z}dz\Big) \tag{A.2}$$

For the left contour integral, we see that we have a pole of order 1 in $z=0$, and poles of order 1 in $z = \pm\frac{1}{r}$. Because $r > 1$, we have $\frac{1}{r} < 1$ and thus the poles $z = \pm\frac{1}{r}$ lie within our contour. The residues are:

$$\begin{aligned} Res &= \lim_{z\to 0} \frac{(z+h)^2(1+hz)^2}{(z-\frac{1}{r})(z+\frac{1}{r})} = 2\pi i\frac{h^2}{-\frac{1}{r^2}} = -r^2h^2 \\ Res(f,\frac{1}{r}) &= \lim_{z\to\frac{1}{r}} \frac{(z+h)^2(1+hz)^2}{z(z+\frac{1}{r})} = \frac{(\frac{1}{r}+h)^2(1+\frac{h}{r})^2}{\frac{1}{r}\frac{2}{r}} \\ Res(f,-\frac{1}{r}) &= \lim_{z\to-\frac{1}{r}} \frac{(z+h)^2(1+hz)^2}{z(z-\frac{1}{r})} = \frac{(-\frac{1}{r}+h)^2(1-\frac{h}{r})^2}{\frac{-1}{r}\frac{-2}{r}} \end{aligned}$$

Applying the residue theorem, and then taking the limit $r \downarrow 1$ we get for the left part of A.2:

$$\begin{aligned} &\lim_{r\downarrow 1}\frac{1}{2\pi i}\oint_{|z|=1} \frac{(z+h)^2(1+hz)^2}{z^2}\frac{z}{(z-\frac{1}{r})(z+\frac{1}{r})}dz \\ &= \lim_{r\downarrow 1}\frac{1}{2\pi i}2\pi i(Res + Res(f,\frac{1}{r}) + Res(f,\frac{-1}{r})) \\ &= \lim_{r\downarrow 1}(-r^2h^2 + \frac{(\frac{1}{r}+h)^2(1+\frac{h}{r})^2}{\frac{1}{r}\frac{2}{r}} + \frac{(-\frac{1}{r}+h)^2(1-\frac{h}{r})^2}{\frac{-1}{r}\frac{-2}{r}} \\ &= -h^2 + \frac{(h+1)^4}{2} + \frac{(h-1)^2(1-h)^2}{2} \\ &= -h^2 + 1 + 6h^2 + h^4 \\ &= h^4 + 5h^2 + 1 \end{aligned}$$

The right contour integral in equation A.2 has pole of order 3 in $z = 0$.That means it has residue

$$\begin{aligned} Res &= \frac{1}{2!}\lim_{z\to 0}\frac{d^2}{dz^2}(z+h)^2(1+hz)^2 \\ &= \frac{1}{2}(2+8h^2+2h^4) = 1 + 4h^2 + h^4 \end{aligned}$$

So the contour integral is equal to

$$\frac{1}{2\pi i}\oint_{|z|=1} \frac{(z+h)^2(1+hz)^2}{z^2}\frac{1}{z}dz = 1 + 4h^2 + h^4$$

Subtracting the contour integrals from each other gives us that

$$\begin{aligned} I_1(x^2) &= \frac{1}{2\pi i}\Big(\oint_{|z|=1} \frac{(z+h)^2(1+hz)^2}{z^2}\frac{z}{(z-\frac{1}{r})(z+\frac{1}{r})}dz - \oint_{|z|=1} \frac{(z+h)^2(1+hz)^2}{z^2}\frac{1}{z}dz\Big) \\ &= h^4 + 5h^2 + 1 - (h^4 + 4h^2 + 1) = h^2 \end{aligned}$$

Recall that $h^2 = c$, we get that $I_1(x^2) = c$, and so the first part of the expectation is $(\kappa - 1)c$. Now for the second part that follows after β. We use again that $|1+hz|^4 = \frac{(z+h)^2(1+hz)^2}{z^2}$so we have

$$\begin{aligned} I_2(x^2) &= \frac{1}{2\pi i}\oint_{|z|=1} |1+hz|^4\frac{1}{z^3}dz \\ &= \frac{1}{2\pi i}\oint_{|z|=1} \frac{(z+h)^2(1+hz)^2}{z^2}\frac{1}{z^3}dz \end{aligned}$$

Now $z = 0$ is a pole of order 5, so the residue is:

$$
\begin{aligned}
Res &= \frac{1}{4!} \lim_{z \to 0} \frac{d^4}{dz^4} (z+h)^2(1+hz)^2 \\
&= \frac{1}{24}(24h^2) \\
&= h^2
\end{aligned}
$$

So the right part of the expectation equals $\beta h^2 = \beta c$. Taking the two parts together, we get that the expectation $\mathbb{E}(X_{x^2}) = (\kappa - 1 + \beta)c$

Variance For the variance, we use the covariance formula from theorem 1.5.1 and recall that $Var(X_f) = Cov(X_f, X_f)$. The formula becomes, again using $f = g = x^2$ and Proposition 1.5.1:

$$
\begin{aligned}
Var(X_{x^2}) &= \kappa J_1(x^2, x^2) + \beta J_2(x^2, x^2) \\
&= \kappa \lim_{r \downarrow 1} J_1(x^2, x^2, r) + \beta J_2(x^2, x^2) \\
&= \kappa \lim_{r \downarrow 1} \frac{-1}{4\pi^2} \oint_{|z_2|=1} \oint_{|z_1|=1} \frac{(|1+hz_1|)^4)(|1+hz_2|^4)}{(z_1 - rz_2)^2} dz_1 dz_2 \\
&+ \beta \frac{-1}{4\pi^2} \oint_{|z_1|=1} \frac{(|1+hz_1|)^4}{z_1^2} dz_1 \oint_{|z_2|=1} \frac{(|1+hz_2|)^4}{z_2^2} dz_2
\end{aligned}
$$

We again first compute the contour integral(s) after κ, $J_1(x^2, x^2, r)$. The inner contour integral becomes, after rewriting $|1+hz_1|^4$

$$
\oint_{|z_1|=1} \frac{(z_1+h)^2(1+hz_1)^2}{z_1^2} \frac{(|1+hz_2|^4)}{(z_1 - rz_2)^2} dz_1
$$

The only pole is $z_1 = 0$ of order 2 and not $z_1 = rz_2$, because for fixed $|z_2| = 1$ and $r > 1$, rz_2 lies outside the contour. It's winding number is thus 0, and it's residue adds nothing to the integral. For the residue at $z_1 = 0$we use the chain rule:

$$
\begin{aligned}
Res &= \lim_{z_1 \to 0} \frac{d}{dz_1} (z_1+h)^2(1+hz_1)^2 \frac{(|1+hz_1|^4)}{(z_1 - rz_2)^2} \\
&= (|1+hz_2|^4) \lim_{z_1 \to 0} \frac{d}{dz_1} \Big(\frac{(z_1+h)(1+hz_1)}{(z_1 - rz_2)} \Big)^2 \\
&= (|1+hz_2|^4) \lim_{z_1 \to 0} \frac{d}{dz_1} t(z_1)^2 \\
&= (|1+hz_2|^4) \lim_{z_1 \to 0} 2t(z_1) \frac{d}{dz} t(z_1)
\end{aligned}
$$

Here $t(z_1) = \frac{(z_1+h)(1+hz_1)}{(z_1-rz_2)}$. Its derivative is $\frac{d}{dz} t(z_1) = \frac{(z_1-rz_2)(1+2hz_1^2+h^2)-(z_1+h)(1+hz_1)}{(z_1-rz_2)^2}$ Taking the limit $z_1 \to 0$ in both the function and the derivative gives us:

$$
\begin{aligned}
\lim_{z_1 \to 0} t(z_1) &= \frac{h}{-rz_2} \\
\lim_{z_1 \to 0} t'(z_1) &= \frac{(-rz_2)(1+h^2) - h}{(-rz_2)^2}
\end{aligned}
$$

So the residue is equal to

$$
2(|1+hz_2|^4) \frac{h}{-rz_2} \frac{(-rz_2)(1+h^2) - h}{(-rz_2)^2}
$$

So the inner contour integral is equal to $2\pi i \cdot 2(|1+hz_2|^4)\frac{h}{-rz_2}\frac{(-rz_2)(1+h^2)-h}{(-rz_2)^2}$ This becomes our new integrand for the outer contour integral where we integrate over z_2. That outer integral becomes

$$\begin{aligned}
&\frac{-1}{4\pi^2}\oint_{|z_2|=1} 2\pi i(|1+hz_2|^4)2\frac{h}{-rz_2}\frac{(-rz_2)(1+h^2)-h}{(-rz_2)^2}dz_2 \\
&=\frac{1}{2\pi i}\oint_{|z_2|=1}\frac{(z_2+h)^2(1+hz_2)^2}{z_2^2}\frac{2h}{-rz_2}\frac{(-rz_2)(1+h^2)-h}{(-rz_2)^2}dz_2 \\
&=\frac{1}{2\pi i}\oint_{|z_2|=1}\frac{(z_2+h)^2(1+hz_2)^2}{z_2^2}\frac{2h}{z_2}\frac{(-rz_2)(1+h^2)-h}{z_2^2}\frac{1}{-r^3}dz_2
\end{aligned}$$

We have a pole of order 5 in $z_2=0$. The residue in $z_2=0$ is

$$\begin{aligned}
Res &= \frac{1}{4!}\lim_{z_2\to 0}\frac{d^4}{dz_2^4}\frac{(z_2+h)^2(1+hz_2)^2(2h)((-rz_2)(1+h^2)-h)}{-r^3} \\
&= \frac{1}{-24r^3}\lim_{z_2\to 0}(-48h^2((5h^3+5h)rz_2+(2h^4+4h^2+2)r+h^2)) \\
&= \frac{2}{r^3}h^2(h^2+r(2h^4+4h^2+2))
\end{aligned}$$

So we get that

$$\frac{1}{2\pi i}\oint_{|z_1|=1}\frac{(z_1+h)^2(1+hz_1)^2}{z_1^2}\frac{2h}{z_1}\frac{z_1(1+h^2)+rh}{z_1^2}dz_1=\frac{2}{r^3}h^2(h^2+r(2h^4+4h^2+2))$$

Taking the limit again with $r\downarrow 1$, we get that

$$\begin{aligned}
&\lim_{r\downarrow 1}\frac{-1}{4\pi^2}\oint_{|z_1|=1}\oint_{|z_2|=1}\frac{(|1+hz_1|)^4)(|1+hz_2|^4)}{(z_1-rz_2)^2}dz_2dz_1 \\
&=\lim_{r\downarrow 1}\frac{2}{r^3}h^2(h^2+r(2h^4+4h^2+2)) \\
&=2(h^2)(h^2+2h^4+4h^2+2) \\
&=2h^2(2h^4+5h^2+2) \\
&=4h^6+10h^4+4h^2
\end{aligned}$$

Now we compute the β part $J_2(x^2,x^2)$ of the variance:

$$\beta\frac{-1}{4\pi^2}\oint_{|z_1|=1}\frac{(|1+hz_1|)^4}{z_1^2}dz_1\oint_{|z_2|=1}\frac{(|1+hz_2|)^4}{z_2^2}dz_2$$

These are 2 loose contour integrals. The first one becomes:

$$\begin{aligned}
&\oint_{|z_1|=1}\frac{(|1+hz_1|)^4}{z_1^2}dz_1 \\
&=\oint_{|z_1|=1}\frac{(z_1+h)^2(1+hz_1)^2}{z_1^2}\frac{1}{z_1^2}dz_1
\end{aligned}$$

It has pole of order 4: so for the residue we get

$$\begin{aligned}
Res &= \frac{1}{3!}\lim_{z_1\to 0}\frac{d^3}{dz_1^3}(z_1+h)^2(1+hz_1)^2 \\
&= \frac{1}{6}(12h+12h^3) \\
&= (2h+2h^3)
\end{aligned}$$

The first contour integral is then equal to $2\pi i(2h+2h^3)$
The second contour integral is

$$\begin{aligned}
&\oint_{|z_2|=1} \frac{(|1+hz_2|)^4}{z_2^2} dz_2 \\
&= \oint_{|z_2|=1} \frac{(z_2+h)^2(1+hz_2)^2}{z_2^2} \frac{1}{z_2^2} dz_2
\end{aligned}$$

This is the same integral with z_2 instead of z_1. So this contour integral is also equal to $2\pi i(2h+2h^3)$. Multiplying the values together gives us the β part of the variance:

$$\begin{aligned}
&\beta \frac{-1}{4\pi^2} \oint_{|z_1|=1} \frac{(|1+hz_1|)^4}{z_1^2} dz_1 \oint_{|z_2|=1} \frac{(|1+hz_2|)^4}{z_2^2} dz \\
&= \beta \frac{-1}{4\pi^2} \big(2\pi i(2h+2h^3) 2\pi i(2h+2h^3)\big) \\
&= \beta(2h+2h^3)^2 \\
&= \beta(4h^2+8h^4+4h^6)
\end{aligned}$$

Now we add together the κ -and β part, $J_1(x^2,x^2)$ and $J_2(x^2,x^2)$ and use $h=\sqrt{c}$ to get

$$\begin{aligned}
Var(X_f) &= \kappa(4h^6+10h^4+4h^2) + \beta(4h^2+8h^4+4h^6) \\
&= 4h^6(\kappa+\beta) + 8h^4(\kappa+\beta) + 4h^2(\kappa+\beta) + 2\kappa h^4 \\
&= 4(\kappa+\beta)(c^3+2c^2+c) + 2\kappa c^2
\end{aligned}$$

Conclusions We have an expectation and variance, but we have to put it into a form such that we can apply it. Recall that $X_{x^2} = p\big\{\frac{1}{p}\sum_{i=1}^p \lambda_i^2 - F_c(x^2)\big\}$. We note that $F_c(x^2) = \int x^2 p_c(x)dx$ is the second moment of the Marcenko Pastur distribution. Using Proposition 1.2.1, we get that

$$\begin{aligned}
F_c(x^2) &= \int x^2 dp_c(x)dx = \mu_2 \\
&= \sum_{r=0}^{1} \frac{1}{r+1} \binom{2}{r}\binom{1}{r} c^r \\
&= \binom{2}{0}\binom{1}{0} c^0 + \frac{1}{2}\binom{2}{1}\binom{1}{1} c^1 = 1+c
\end{aligned}$$

In other words:

$$\begin{aligned}
X_{x^2} &= p\big\{\frac{1}{p}\sum_{i=1}^p \lambda_i^2 - F_c(x^2)\big\} \\
&= \sum_{i=1}^p \lambda_i^2 - p(1+c) \\
&= ||S||_F^2 - p(1+c) \xrightarrow{D} N(\mu_{x^2}, \sigma_{x^2}^2)
\end{aligned}$$

So the squared Frobenius norm minus $p(1+c)$ is normal in the limit with expectation $\mu_{x^2} = (\kappa-1+\beta)c$ and variance $\sigma_{x^2}^2 = 4(\kappa+\beta)(c^3+2c^2+c)+2\kappa c^2$

A.1.2 Limiting distribution of $tr(S_n)$

We proceed similarly for $tr(S) = \sum_{i=1}^p \lambda_i$. The trace of S_n is a linear spectral statistic with $f=x$. Namely:

$$tr(S_n) = \sum_{i=1}^p \lambda_i$$

Using again the Central Limit Theorem 1.5.1, we get that under H_0:
$X_n(x) = p\{\frac{1}{p}\sum_{i=1}^{p} \lambda_i - F_c(x)\}$ converges to a normal variable.
But $p(\frac{1}{p}\sum_{i=1}^{p} \lambda_i) = \sum_{i=1}^{p} \lambda_i$, which is our trace. Our goal is then to find the expectation and variance of $tr(S) - pF_c(x)$

Expectation For the expectation we get again if we apply the Central Limit Theorem 1.5.1 and Proposition 1.5.1 with $f = x$

$$\begin{aligned}\mathbb{E}(X_x) &= (\kappa - 1)I_1(x) + \beta I_2(x) \\ &= (\kappa - 1)\lim_{r\downarrow 1} I_1(x, r) + \beta I_2(x) \\ &= (\kappa - 1)\lim_{r\downarrow 1} \frac{1}{2\pi i}\oint_{|z|=1} |1 + hz|^2 \big[\frac{z}{z^2 - r^{-2}} - \frac{1}{z}\big]dz + \beta\frac{1}{2\pi i}\oint_{|z|=1} |1 + hz|^2 \frac{1}{z^3}dz\end{aligned}$$

We first compute the integral that follows $\kappa - 1, I_1(x)$. Using our remark about $|1 + hz|^2$ from earlier and splitting it up gives us

$$I_1(x) = \lim_{r\downarrow 1} I_1(x, r) = \frac{1}{2\pi i}\lim_{r\downarrow 1}\oint_{|z|=1} |1 + hz|^2 \big[\frac{z}{z^2 - r^{-2}} - \frac{1}{z}\big]dz \tag{A.3}$$

$$= \frac{1}{2\pi i}\lim_{r\downarrow 1}\Big\{\oint_{|z|=1} \frac{(z + h)(1 + hz)}{z}\frac{z}{(z - \frac{1}{r})(z + \frac{1}{r})}dz - \oint_{|z|=1} \frac{(z + h)(1 + hz)}{z}\frac{1}{z}dz\Big\} \tag{A.4}$$

The left integral has only poles at $z = \pm\frac{1}{r}$ this time, and has residues:

$$\begin{aligned}Res(f, \frac{1}{r}) &= \lim_{z\to\frac{1}{r}} \frac{(z + h)(1 + hz)}{(z + \frac{1}{r})} = \frac{(\frac{1}{r} + h)(1 + \frac{h}{r})}{\frac{2}{r}} \\ Res(f, -\frac{1}{r}) &= \lim_{z\to -\frac{1}{r}} \frac{(z + h)(1 + hz)}{(z - \frac{1}{r})} = \frac{(-\frac{1}{r} + h)(1 - \frac{h}{r})}{-\frac{2}{r}}\end{aligned}$$

Adding the residues together gives us for the first integral

$$\begin{aligned}&\frac{(\frac{1}{r} + h)(1 + \frac{h}{r})}{\frac{2}{r}} + \frac{(-\frac{1}{r} + h)(1 - \frac{h}{r})}{-\frac{2}{r}} \\ &= \frac{(1 + rh)(1 + \frac{h}{r})}{2} + \frac{(1 - rh)(1 - \frac{h}{r})}{2} \\ &= \frac{1}{2}(1 + \frac{h}{r} + rh + h^2 + 1 - \frac{h}{r} - rh + h^2) \\ &= \frac{1}{2}(2 + 2h^2) \\ &= 1 + h^2\end{aligned}$$

For the second integral we have:

$$\frac{1}{2\pi i}\oint_{|z|=1} \frac{(z + h)(1 + hz)}{z}\frac{1}{z}dz$$

This one has pole of order 2 at $z = 0$. The residue is then:

$$\begin{aligned}Res &= \lim_{z\to 0} \frac{d}{dz}(z + h)(1 + hz) \\ &= \lim_{z\to 0} \frac{d}{dz}(z + hz^2 + h + h^2 z) \\ &= \lim_{z\to 0}(1 + 2hz + h^2) \\ &= 1 + h^2\end{aligned}$$

The contour integrals are equal. Since we subtract the second one from the first, we get that $I_1(x)$ is 0.

For the β part, $I_2(x)$:

$$\begin{aligned} I_2(x) &= \frac{1}{2\pi i}\oint_{|z|=1} f(|1+hz|^2)\frac{1}{z^3}dz \\ &= \frac{1}{2\pi i}\oint_{|z|=1} \frac{(z+h)(1+hz)}{z}\frac{1}{z^3}dz \end{aligned}$$

This has pole of order 4 at $z = 0$, and residue:

$$Res = \frac{1}{3!}\lim_{z\to 0}\frac{d^3}{dz^3}(z+h)(1+hz)$$

However, the polynomial being differentiated has highest power of hz^2, and so the whole thing vanishes when differentiating 3 times. Thus the residue equals 0, and so the whole contour integral $I_2(x)$ is 0.

Since both $I_1(x)$ and $I_2(x)$ are 0, the expectation $\mathbb{E}(tr(S) - p)$ is 0.

Variance For the variance we again get:

$$\begin{aligned} Var(X_x) &= \kappa J_1(x,x) + \beta J_2(x,x) \\ &= \kappa \lim_{r\downarrow 1} J_1(x,x,r) + \beta J_2(x,x) \\ &= \kappa \lim_{r\downarrow 1}\frac{-1}{4\pi^2}\oint_{|z_1|=1}\oint_{|z_2|=1}\frac{(|1+hz_1|)^2)(|1+hz_2|^2)}{(z_1-rz_2)^2}dz_2dz_1 \\ &+ \beta\frac{-1}{4\pi^2}\oint_{|z_1|=1}\frac{(|1+hz_1|)^2}{z_1^2}dz_1\oint_{|z_2|=1}\frac{(|1+hz_2|)^2}{z_2^2}dz_2 \end{aligned}$$

We first compute $\lim_{r\downarrow 1} J_1(x,x,r)$ This is:

$$\begin{aligned} \lim_{r\downarrow 1} J_1(x,x,r) &= \lim_{r\downarrow 1}\frac{-1}{4\pi^2}\oint_{|z_2|=1}\oint_{|z_1|=1}\frac{(|1+hz_1|)^2)(|1+hz_2|^2)}{(z_1-rz_2)^2}dz_1dz_2 \\ &= \lim_{r\downarrow 1}\frac{-1}{4\pi^2}\oint_{|z_2|=1}\oint_{|z_1|=1}\frac{(z_1+h)(1+hz_1)}{z_1}\frac{(z_2+h)(1+hz_2)}{z_2}\frac{1}{(z_1-rz_2)^2}dz_1dz_2 \end{aligned}$$

The inner contour integral has pole of order 1 at $z_1 = 0$, and residue:

$$\begin{aligned} Res &= \lim_{z_1\to 0}\frac{(z_2+h)(1+hz_2)}{z_2}(z_1+h)(1+hz_1)\frac{1}{(z_1-rz_2)^2} \\ &= \frac{(z_2+h)(1+hz_2)}{z_2}\frac{h}{(-rz_2)^2} \\ &= \frac{(z_2+h)(1+hz_2)}{z_2^3}\frac{h}{r^2} \end{aligned}$$

So $2\pi i\frac{(z_2+h)(1+hz_2)}{z_2^3}\frac{h}{r^2}$ becomes the integrand in the contour integral over z_2:

$$\begin{aligned} J_1(x,x) &= \lim_{r\downarrow 1} J_1(x,x,r) \\ &= \lim_{r\downarrow 1}\frac{1}{2\pi i}\oint_{|z_2|=1}\frac{(z_2+h)(1+hz_2)}{z_2^3}\frac{h}{r^2}dz_2 \end{aligned}$$

This has pole of order 3 in $z_2 = 0$ with residue:

$$\begin{aligned}
Res &= \frac{1}{2!} \lim_{z_2 \to 0} \frac{d^2}{dz_2^2} (z_2 + h)(1 + hz_2) \frac{h}{r^2} \\
&= \frac{1}{2} \lim_{z_2 \to 0} (2h) \frac{h}{r^2} \\
&= \frac{h^2}{r^2}
\end{aligned}$$

Taking the limit as $\lim_{r \downarrow 1}$, we get that $J_1(x, x) = h^2 = c$.
Now for the β part:

$$J_2(x, x) = \frac{-1}{4\pi^2} \oint_{|z_1|=1} \frac{(|1 + hz_1|)^2}{z_1^2} dz_1 \oint_{|z_2|=1} \frac{(|1 + hz_2|)^2}{z_2^2} dz_2$$

These are the same integrals, just over different variables. We compute one:

$$\begin{aligned}
&\oint_{|z_1|=1} \frac{(|1 + hz_1|)^2}{z_1^2} dz_1 \\
&= \oint_{|z_1|=1} \frac{(z_1 + h)(1 + hz_1)}{z_1} \frac{1}{z_1^2} dz_1
\end{aligned}$$

This has pole of order 3 at $z_1 = 0$, and so residue

$$\begin{aligned}
Res &= \frac{1}{2!} \lim_{z_1 \to 0} \frac{d^2}{dz_1^2} (z_1 + h)(1 + hz_1) \\
&= \frac{1}{2} \lim_{z_1 \to 0} (2h) \\
&= h
\end{aligned}$$

So one contour integral equals $2\pi i h$, and squared it is $\frac{-4\pi^2}{h^2}$. This gives us that $J_2(x, x) = h^2 = c$, just like $J_1(x, x)$
Combining everything gives us:

$$\begin{aligned}
Var(X_x) &= \kappa J_1(x, x) + \beta J_2(x, x) \\
&= \kappa c + \beta c = (\kappa + \beta)c
\end{aligned}$$

Conclusions Again we have an expectation and variance, but we have to put it into a form such that we can apply it. Recall that $X_x = p\big\{\frac{1}{p}\sum_{i=1}^{p} \lambda_i - F_c(x)\big\}$. Here we have that $F_c(x)$ is the expectation of the Marcenko Pastur law, and is 1. In other words:

$$\begin{aligned}
X_x &= p\{\frac{1}{p}\sum_{i=1}^{p} \lambda_i - F_c(x)\} \\
&= \sum_{i=1}^{p} \lambda_i - p \\
&= tr(S_n) - p \xrightarrow{D} N(\mu_x, \sigma_x^2)
\end{aligned}$$

So the trace of S_n minus p is normal in the limit with expectation $\mu_x = 0$ and variance $\sigma_x^2 = (\kappa+\beta)c$

A.1.3 Covariance of $||S_n||_F^2$ and $tr(S_n)$

If you want to find the distribution of $p\hat{\alpha}$ you will need the covariance between the squared Frobenius norm and the trace. In this section we will derive their covariance.
We use again that the squared Frobenius norm and trace are linear spectral statistics with $f = x^2$ and $g = x$ respectively. We can apply the Central Limit Theorem 1.5.1 and Proposition 1.5.1 to

find a formula for their covariance:

$$\begin{aligned}
Cov(X_{x^2}, X_x) &= \kappa J_1(x^2, x) + \beta J_2(x^2, x) \\
&= \kappa \lim_{r\downarrow 1} J_1(x^2, x, r) + \beta J_2(x^2, x) \\
&= \kappa \lim_{r\downarrow 1} -\frac{1}{4\pi^2} \oint_{|z_2|=1} \oint_{|z_1|=1} \frac{(|1+hz_1|^4)(|1+hz_2|^2)}{(z_1 - rz_2)^2} dz_1 dz_2 \\
&+ \beta \frac{-1}{4\pi^2} \oint_{|z_1|=1} \frac{(|1+hz_1|^4)}{z_1^2} dz_1 \oint_{|z_2|=1} \frac{(|1+hz_2|^2)}{z_2^2} dz_2
\end{aligned}$$

We first compute $J_1(x^2, x, r)$ This is

$$\begin{aligned}
J_1(x^2, x, r) &= \lim_{r\downarrow 1} -\frac{1}{4\pi^2} \oint_{|z_2|=1} \oint_{|z_1|=1} \frac{(|1+hz_1|^4)(|1+hz_2|^2)}{(z_1 - rz_2)^2} dz_1 dz_2 \\
&= \lim_{r\downarrow 1} -\frac{1}{4\pi^2} \oint_{|z_2|=1} \oint_{|z_1|=1} \frac{(z_1+h)^2(1+hz_1)^2}{z_1^2} \frac{(z_2+h)(1+hz_2)}{z_2} \frac{1}{(z_1 - rz_2)^2} dz_1 dz_2
\end{aligned}$$

The inner contour integral has pole of order 2 at $z_1 = 0$, because rz_2 lies outside the contour. That means it has residue:

$$\begin{aligned}
Res &= \lim_{z_1\to 0} \frac{d}{dz_1} (z_1+h)^2(1+hz_1)^2 \frac{(z_2+h)(1+hz_2)}{z_2} \frac{1}{(z_1 - rz_2)^2} \\
&= \frac{(z_2+h)(1+hz_2)}{z_2} \lim_{z_1\to 0} \frac{d}{dz_1} q(z_1)^2 \\
&= \frac{(z_2+h)(1+hz_2)}{z_2} \lim_{z_1\to 0} 2q(z_1) \frac{d}{dz_1} q(z_1)
\end{aligned}$$

Here $q(z_1) = \frac{(z_1+h)(1+hz_1)}{(z_1-rz_2)}$, with derivative $\frac{d}{dz_1} q(z_1) = \frac{(z_1-rz_2)(1+2hz_1+h^2)-(z_1+h)(1+hz_1)}{(z_1-rz_2)^2}$ The limits are:

$$\begin{aligned}
\lim_{z_1\to 0} q(z_1) &= \frac{h}{-rz_2} \\
\lim_{z_1\to 0} q'(z_1) &= \frac{-rz_2(1+h^2) - h}{r^2 z_2^2}
\end{aligned}$$

The residue then becomes:

$$Res = \frac{(z_2+h)(1+hz_2)}{z_2} 2 \frac{h}{-rz_2} \frac{-rz_2(1+h^2) - h}{r^2 z_2^2}$$

This times $2\pi i$ becomes the integrand for the outer contour integral, which becomes

$$J_1(x^2, x, r) = \lim_{r\downarrow 1} = \frac{1}{2\pi i} \oint_{|z_2|=1} \frac{(z_2+h)(1+hz_2)}{z_2} \frac{2h}{-rz_2} \frac{-rz_2(1+h^2) - h}{r^2 z_2^2} dz_2$$

We have a pole of order 4 in $z_2 = 0$, and so residue:

$$\begin{aligned}
Res &= \frac{1}{3!} \frac{1}{-r^3} \lim_{z_2\to 0} \frac{d^3}{dz_2^3} (z_2+h)(1+hz_2)(2h)(-rz_2(1+h^2) - h) \\
&= \frac{1}{6} \frac{2}{-r^3} \lim_{z_2\to 0} -6h^2(h^2+1)r \\
&= 2rh^2(h^2+1)
\end{aligned}$$

Taking the limit as $r \downarrow 1$ gives us that $J_1(x^2, x) = 2h^2(h^2+1) = 2c^2 + 2c$
Now we compute $J_2(x^2, x)$:

$$\begin{aligned} J_2(x^2,x) &= \frac{-1}{4\pi^2} \oint_{|z_1|=1} \frac{(|1+hz_1|^4)}{z_1^2} dz_1 \oint_{|z_2|=1} \frac{(|1+hz_2|^2)}{z_2^2} dz_2 \\ &= \frac{-1}{4\pi^2} \oint_{|z_1|=1} \frac{(z_1+h)^2(1+hz_1)^2}{z_1^2} \frac{1}{z_1^2} dz_1 \oint_{|z_2|=1} \frac{(z_1+h)(1+hz_1)}{z_1} \frac{1}{z_2^2} dz_2 \end{aligned}$$

The left contour integral has pole of order 4 in $z_1 = 0$ and so it has residue

$$\begin{aligned} Res &= \frac{1}{3!} \lim_{z_1 \to 0} \frac{d^3}{dz_1^3} (z_1+h)^2(1+hz_1)^2 \\ &= \frac{1}{6} \lim_{z_1 \to 0} 12h(2hz_1 + h^2 + 1) \\ &= 2h(h^2+1) \end{aligned}$$

The right contour integral has pole of order 3 in $z_2 = 0$ and so it has residue

$$\begin{aligned} Res &= \frac{1}{2!} \lim_{z_2 \to 0} \frac{d^2}{dz_2^2} (z_1+h)(1+hz_1) \\ &= \frac{1}{2} \lim_{z_2 \to 0} 2h \\ &= h \end{aligned}$$

Multiplying the contour integrals together gives us that

$$J_2(x^2,x) = 2h(h^2+1)h = 2h^2(h^2+1) = 2c^2 + 2c$$

Combining the results for $J_1(x^2,x), J_2(x^2,x)$ gives us

$$Cov(X_{x^2}, X_x) = \kappa(2c^2+2c) + \beta(2c^2+2c) = 2(\kappa+\beta)(c^2+c)$$

A.2 Derivation limiting distribution $\hat{\alpha}^*$

A.2.1 Joint Normal Distribution of $||S||_F^2$ and $tr(S_n)^2$

Before we can move on to the derivation of $\hat{\alpha}^*$, we need to observe the joint normality of the Frobenius norm and the trace. Since they have a covariance $2(\kappa+\beta)(c^2+c)$, which is nonzero, they are dependent. However they can still be jointly normal. It is however important that we subtract $p(1+c)$ from the Frobenius norm and p from the trace, otherwise these quantities would become unbounded as $p \to \infty$. We then get

$$(||S_n||_F^2 - p(1+c), tr(S_n) - p)^T \xrightarrow{D} N(\mu, \mathbf{\Sigma})$$

where $\mu = ((\kappa - 1 + \beta)c, 0)) =^T$ and $\Sigma = \begin{pmatrix} 4(\kappa+\beta)(c^3+2c^2+c) + 2\kappa c^2) & 2(\kappa+\beta)(c^2+c) \\ 2(\kappa+\beta)(c^2+c) & (\kappa+\beta)c \end{pmatrix}$

A.2.2 Delta method applied to $\hat{\alpha}^*$

We have

$$\hat{\alpha}^* = 1 - \frac{c \cdot tr(S_n)^2}{p||S_n||_F^2 - tr(S_n)^2}$$

But also we have that:

$$\begin{Bmatrix} ||S_n||_F^2 - p(1+c) \\ tr(S_n) - p \end{Bmatrix} = p \begin{Bmatrix} \frac{1}{p}\sum_{i=1}^p \lambda_i^2 - (1+c) \\ \frac{1}{p}\sum_{i=1}^p \lambda_i - 1 \end{Bmatrix} \xrightarrow{D} \begin{bmatrix} \mathbf{X_1} \\ \mathbf{X_2} \end{bmatrix}$$

$$\begin{aligned}
\mathbf{X_1} &\sim N((\kappa-1+\beta)c, 4(\kappa+\beta)(c^3+2c^2+c)+2\kappa c^2) \\
\mathbf{X_2} &\sim N((0,(\kappa+\beta)c) \\
Cov(\mathbf{X_1},\mathbf{X_2}) &= 2(\kappa+\beta)(c^2+c)
\end{aligned}$$

If we can write $\hat{\alpha}^*$ as a function of $\frac{1}{p}\sum_{i=1}^{p}\lambda_i^2$ and $\frac{1}{p}\sum_{i=1}^{p}\lambda_i$, we can apply the Delta method with $\mathbf{a}=\begin{pmatrix}1+c\\1\end{pmatrix}$, $p\to\infty$ and $b=1$
To this end, let us rewrite

$$\begin{aligned}
\hat{\alpha}^* &= 1-\frac{c\cdot tr(S_n)^2}{p||S_n||_F^2-(tr(S_n))^2} \\
&= 1-\frac{c\cdot p^2(\frac{1}{p}tr(S_n))^2}{p^2\frac{1}{p}||S_n||_F^2-p^2(\frac{1}{p}tr(S_n))^2} \\
&= 1-\frac{cy^2}{x-y^2}
\end{aligned}$$

The last step is justified since we can divide away p^2, and substitute $x=\frac{1}{p}||S_n||_F^2$ and $y=\frac{1}{p}tr(S_n)$. We substitute for simplicity.
We now observe that our $\hat{\alpha}^*$ is a function $g:\mathbb{R}^2\to\mathbb{R}$ for the sake of the Delta method: Namely

$$\hat{\alpha}^*=g(x,y)=1-\frac{cy^2}{x-y^2}$$

The Delta method now implies that

$$p\big\{\hat{\alpha}^*-g(\mathbf{a})\big\}\xrightarrow{D}[\nabla g(\mathbf{a})]^T\begin{bmatrix}\mathbf{X_1}\\\mathbf{X_2}\end{bmatrix},\mathbf{a}=\begin{pmatrix}1+c\\1\end{pmatrix}$$

We compute then:

$$\begin{aligned}
g(\mathbf{a}) &= g(1+c,1) \\
&= 1-\frac{c(1^2)}{1+c-1^2} \\
&= 1-\frac{c}{c}=0
\end{aligned}$$

For $[\nabla g(\mathbf{a})]^T=[\frac{\partial g}{\partial x}(\mathbf{a}),\frac{\partial g}{\partial y}(\mathbf{a})]$, we compute the partial derivatives with respect to x and y:

$$\begin{aligned}
\frac{\partial g}{\partial x} &= \frac{\partial}{\partial x}\Big\{1-\frac{cy^2}{x-y^2}\Big\} \\
&= \frac{cy^2}{(x-y^2)^2} \\
\frac{\partial g}{\partial y} &= \frac{\partial}{\partial y}\Big\{1-\frac{cy^2}{x-y^2}\Big\} \\
&= -\frac{2cxy}{(x-y^2)^2}
\end{aligned}$$

If we then put in $\mathbf{a}=\begin{pmatrix}1+c\\1\end{pmatrix}$, we get for the partial derivatives:

$$\begin{aligned}
\frac{\partial g}{\partial x}(\mathbf{a}) &= \frac{c(1^2)}{(1+c-(1^2))^2} \\
&= \frac{c}{c^2}=\frac{1}{c} \\
\frac{\partial g}{\partial y}(\mathbf{a}) &= -\frac{2c(1+c)\cdot 1}{(1+c-(1^2))^2} \\
&= -\frac{2c(1+c)}{c^2}=-\frac{2(1+c)}{c}
\end{aligned}$$

So we get that

$$[\nabla g(\mathbf{a})]^T = [\frac{1}{c}, \frac{2(1+c)}{c}]$$

Combining this all implies that

$$\begin{aligned}
p\hat{\alpha}^* &\xrightarrow{D} [\frac{1}{c}, \frac{2(1+c)}{c}] \begin{bmatrix} \mathbf{X_1} \\ \mathbf{X_2} \end{bmatrix} \\
&= \frac{1}{c}\mathbf{X_1} - \frac{2(1+c)}{c}\mathbf{X_2}
\end{aligned}$$

So $p\hat{\alpha}^*$ is the sum of two normal variables in the limit, and is thus also normal! It has expectation

$$\begin{aligned}
\mathbb{E}(p\hat{\alpha}^*) &= \mathbb{E}\Big(\frac{1}{c}\mathbf{X_1} - \frac{2(1+c)}{c}\mathbf{X_2}\Big) \\
&= \frac{1}{c}\mathbb{E}(\mathbf{X_1}) - \frac{2(1+c)}{c}\mathbb{E}(\mathbf{X_2}) \\
&= \frac{1}{c}(\kappa - 1 + \beta)c - \frac{2(1+c)}{c} \cdot 0 \\
&= \kappa - 1 + \beta
\end{aligned}$$

It has variance

$$\begin{aligned}
Var(p\hat{\alpha}^*) &= Var\Big(\frac{1}{c}\mathbf{X_1} - \frac{2(1+c)}{c}\mathbf{X_2}\Big) \\
&= \frac{1}{c^2}Var(\mathbf{X_1}) + (-\frac{2(1+c)}{c})^2 Var(\mathbf{X_2}) - 2 \cdot \frac{1}{c}\frac{2(1+c)}{c}Cov(\mathbf{X_1}, \mathbf{X_2}) \\
&= \frac{1}{c^2}(4(\kappa+\beta)(c^3+2c^2+c) + 2\kappa c^2) + (\frac{2(1+c)}{c})^2(\kappa+\beta)(c) - 2\frac{1}{c}\frac{2(1+c)}{c}2(\kappa+\beta)(c^2+c) \\
&= 2\kappa
\end{aligned}$$

Thus

$$p\hat{\alpha}^* \xrightarrow{D} N(\kappa - 1 + \beta, 2\kappa)$$

We now have derived a ready to use statistic in the high-dimensional case based on a estimator other than the regular sample covariance matrix.

Appendix B

Matlab Code

```
function y = TraceNorm(S);
y = trace((S*transpose(S))^(1/2));
end

function y = Frob(S);
y = sqrt(trace(S*transpose(S)));
end
```

B.1 Codes used for sizes / power curves

B.1.1 Empirical size

```
%% EMPIRICAL SIZES
clearvars;
tic

A = 1e2;

pval =[8,16,32,64,96,126,192,224,256];
n = 256;
w = 1.6449;
kappa =2;
beta = 1.5; %% OR 0 IF NORMAL DATA

powervecCLRT = zeros(length(pval),1);
powervecCJ = zeros(length(pval),1);
powervecLS =zeros(length(pval),1);
for j = 1:length(pval)
    p = pval(j);
    c = p/n;
    Sigma_0 = eye(p);
    countCJ = 0;
    countCLRT =0;
    countLS = 0;
    for i = 1:A
        Y = gamrnd(4,1/2,p,n)-2;
        %Y = randn(p,n);
        S = 1/n*(Y*transpose(Y));
        LambdaN = p*log(trace(S)/p)-log(det(S)); %
        if (LambdaN+(p-n)*log(1-c)-p)>w*sqrt(-kappa*log(1-c)-kappa*c)+ (-(kappa-1)*log
            countCLRT = countCLRT+1;
        end
        %CJ TEST
        T2 =(p^2*n/2)*trace((S/trace(S)-eye(p)/p)^2);
        U = (2/(n*p))*T2;
        if (n*U-p) > w*sqrt(2*kappa)+(kappa-1+beta);
            countCJ = countCJ + 1;
        end

        %LINEAR SHRINKAGE
        alphahat = 1-(1/n*trace(S)^2*Frob(Sigma_0)^2)/(Frob(S)^2*Frob(Sigma_0)^2-trace
        if p*alphahat > w*sqrt(2*kappa)+(kappa-1+beta);
            countLS = countLS+1;
        end
    end
    powervecCLRT(j) = countCLRT/A;
    powervecCJ(j) = countCJ/A;
    powervecLS(j) = countLS/A;
end
toc
```

B.1.2 Power plots

```
clearvars;
p = 64;
n = 128;
c=p/n;
rho = -0.3:0.01:0.3;

kappa = 2;
beta = 0;

A = 1e3;
powervecCLRT = zeros(length(rho),1);
powervecCJ = zeros(length(rho),1);
powervecLS =zeros(length(rho),1);

alpha = 0:0.01:1;

$Sigma_0$ = eye(p);
tic
for j = 1:length(rho)
    r = rho(j);
    Sigma = (1-r)*eye(p) + r*ones(p);
    countLS = 0;
    countCLRT = 0;
    countCJ = 0;

    for i = 1:A;
        X = randn(p,n);
        Y = Sigma^(1/2)*X;
        S = 1/n*(Y*transpose(Y));
        %statistics
        %CLRT
        LambdaN = p*log(trace(S)/p)-log(det(S));
        if (LambdaN+(p-n)*log(1-c)-p)>1.645*sqrt(-kappa*log(1-c)-kappa*c)+ (-(kappa-1
            countCLRT = countCLRT+1;
        end

        %CJ TEST
        T2 =(p^2*n/2)*trace((S/trace(S)-eye(p)/p)^2);
        U = (2/(n*p))*T2;
        if (n*U-p) > 1.645*sqrt(2*kappa)+(kappa-1+beta);
            countCJ = countCJ + 1;
        end

        %LINEAR SHRINKAGE
        alphahat = 1-(1/n*trace(S)^2*Frob(Sigma_0)^2)/(Frob(S)^2*Frob(Sigma_0)^2-trace
        if p*alphahat > 1.645*sqrt(2*kappa)+(kappa-1+beta);
            countLS = countLS+1;
        end

    end
    powervecCLRT(j) = countCLRT/A;
    powervecCJ(j) = countCJ/A;
    powervecLS(j) = countLS/A;
end
toc
figure
plot(rho,powervecCLRT);
hold on
plot(rho,powervecCJ);
hold on
plot(rho,powervecLS);
hold off
xlabel('Rho')
ylabel('Power')
legend('CLRT','CJ','LS');
title(['c=',num2str(c)]);
```

B.1.3 ROC plots

```
%% ROC CURVES
clearvars;
p = 112;
n = 128;
c=p/n;
gamma = 0.1;
rho = 0.006;
delta = 0.05;

alpha = 0:0.01:1;
x = sqrt(2)*erfinv(1-2*alpha);

kappa = 2;
beta = 0;
A = 1e4;

powervecCLRT = zeros(length(x),1);
powervecCJ = zeros(length(x),1);
powervecLS =zeros(length(x),1);
Sigma_0 = eye(p);
Sigma = (1-rho)*eye(p)+rho*ones(p);
%Sigma = zeros(p);

% d = delta;
%     for a = 1:p;
%         for b = 1:p;
%             Sigma(a,b) = d^(abs(a-b));
%           end
%       end
r = 1/2;
lendiag = floor(r*p);
Sigma = diag([ones(1,p-lendiag),(1+gamma)*ones(1,lendiag)]);
tic
for j = 1:length(x);

    w = x(j);

    countLS = 0;
    countCLRT = 0;
    countCJ = 0;
    for i = 1:A;
        X = randn(p,n);
        Y = Sigma^(1/2)*X;
        S = 1/n*(Y*transpose(Y));
        %statistics
        %CLRT
        LambdaN = p*log(trace(S)/p)-log(det(S)); %
        if (LambdaN+(p-n)*log(1-c)-p)>w*sqrt(-kappa*log(1-c)-kappa*c)+ (-(kappa-1)*log
            countCLRT = countCLRT+1;
        end

        %CJ TEST
        T2 =(p^2*n/2)*trace((S/trace(S)-eye(p)/p)^2);
        U = (2/(n*p))*T2;
        if (n*U-p) > w*sqrt(2*kappa)+(kappa-1+beta);
            countCJ = countCJ + 1;
        end

        %LINEAR SHRINKAGE
        alphahat = 1-(1/n*trace(S)^2*Frob(Sigma_0)^2)/(Frob(S)^2*Frob(Sigma_0)^2-trace
        if p*alphahat > w*sqrt(2*kappa)+(kappa-1+beta);
            countLS = countLS+1;
        end
    end
    powervecCLRT(j) = countCLRT/A;
    powervecCJ(j) = countCJ/A;
    powervecLS(j) = countLS/A;
end
toc
figure
plot(alpha,powervecCLRT);
hold on
plot(alpha,powervecCJ);
hold on
plot(alpha,powervecLS);
hold on
plot(alpha,alpha);
hold off
xlabel('False␣positive␣rate␣alpha')
ylabel('True␣positive␣rate')
legend('CLRT','CJ','LS','Standard');
%title(['Delta = ',num2str(delta),', c =',num2str(c)]);
title(['Gamma␣=␣',num2str(gamma),',␣c␣=',num2str(c),',␣ratio␣=␣',num2str(r)]);
```

Appendix C

List of mathematical symbols

Symbol	Meaning
n	Sample size
p	Dimension
c	ratio p/n
S_n	Sample covariance matrix
λ_i	i-th sample eigenvalue of S_n
Σ_n	True underlying population covariance matrix
Σ_0	Prior belief for the covariance matrix
h	$\sqrt{c}$
κ	Indicator of nature of data, $\kappa = 2$ if our data is real and $\kappa = 1$ if our data is complex
β	Fourth moment - κ -1
H_0	Null hypothesis, $H_0 : \Sigma_n = I$
H_1	Alternative hypothesis one wants to test for
$\|\|S\|\|_F^2$	Frobenius norm of a matrix S
$tr(S)$	Trace of a matrix S

Appendix D

Bibliography

Bibliography

[1] Bodnar, Gupta, Parolya. (2014) *On the strong convergence of the optimal linear shrinkage estimator for large dimensional covariance matrix*

[2] Yao, Zheng, Bai. (2015) *Large Sample Covariance Matrices and High-Dimensional Data Analysis*

[3] Wang, Yao. (2013) *On the sphericity test with large-dimensional observations*

[4] Freitag, Busam (2005) *Complex Analysis*

[5] Fraleigh, Beauregard (1995) *Linear algebra, 3rd edition*

[6] Sadun (200) *Applied Linear Algebra, The Decoupling Principle*

[7] J.L. Doob (1932) *The limiting distributions of certain statistics*

[8] T.W. Anderson (1984) *An Introduction to Multivariate Statistical Analysis*

[9] S. John (1971) *Some optimal multivariate tests*

[10] Grimmett, Welsh (1986) *Probability, an introduction*

[11] Ledoit, Peché (2011) Eigenvectors of some large sample covariance matrix ensembles

[12] Ledoit, Wolf(2011) Nonlinear shrinkage estimation of large-dimensional covariance matrices

[13] Marčenko, V.A., Pastur, L.A.: *Distribution of eigenvalues for some sets of random matrices. Math. USSR-Sb. 1, 457–486* (1967)

[14] Silverstein *Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. J. Multivariate Anal. 55, 331–339* (1995)

[15] Choi, S.I., Silverstein, J.W.: *Analysis of the limiting spectral distribution of large dimensional random matrices. J. Multivariate Anal. 54, 295–309* (1995)

[16] Li, Aue, Paul: (2018) *High-dimensional general linear hypothesis tests via non-linear spectral shrinkage*

[17] Onatski, Moreira, Hallin (2013) *Asymptotic power of sphericity tests for high-dimensional data*