Likelihood Ratio Test for Sphericity under a Multivariate Normal Distribution assumption

Van Thuan Romoli

MT4599: Project in Mathematics / Statistics

I certify that this project report has been written by me, is a record of work carried out by me, and is essentially different from work undertaken for any other purpose or assessment.

University of St Andrews

School of Mathematics and Statistics

Under the supervision of
Dr Regina M. B. Bispo and Dr Filipe J. Marques

Dedication

To Cristina, Claudio, and Thuy. La mia famiglia. I would like to thank them for all the sacrifices they have made to support me, and for all of their love.

Acknowledgements

I would like to thank my supervisor, Dr Regina M. B. Bispo, for all her guidance and encouragement throughout the last year of my studies. I would also like to thank Dr Filipe J. Marques for the (many) suggestions and advice received for my dissertation. I appreciate immensely the time spent working together and the effort put into helping me produce my first piece of mathematical writing. I have enjoyed very much the study and research at the University of St Andrews and I am grateful to all the faculty and the staff in the School of Mathematics and Statistics for all their support.

Likelihood Ratio Test for Sphericity under a Multivariate Normal Distribution assumption

Van Thuan Romoli

Abstract

In multivariate statistics, it is often important to assume independence and homoscedasticity. Under the multivariate normal assumption, testing for these conditions corresponds to test sphericity. While, in general, likelihood ratio tests have great properties, the sphericity test is known to be ineffective in detecting departures from sphericity, and to lose accuracy when the assumption of multivariate normality is violated. The exact distribution of the statistic is difficult to compute and consequently, several approximations have been proposed, the most common of which relies on Wilks' theorem to model the negative logarithm of the likelihood ratio test statistic U, using a chi-squared distribution. Alternatively, a second statistic, U', may be used by scaling U with an appropriate constant. This study investigates the properties of the sphericity test using these two statistics. Through extensive simulations, we assess the accuracy and limitations of the sphericity test, providing insight into when the test can be reliably applied in practice. We conclude suggesting on the use of the statistic U', instead of U, since its distribution shows a better approximation to the expected asymptotic behaviour and it allows to achieve a higher power, without increasing the complexity.

Contents

Chapter 1: Introduction

In many statistical methods, it is important to ensure the independence of variables and the homogeneity of variances. Under the multivariate normal assumption, the validation of these assumptions is equivalent to test if the covariance matrix of a random vector is proportional to the identity matrix, which is more commonly called the sphericity structure. This name is due to the fact that data extracted from a population of independent variables with equal variances tend to have a spherical graphical representation. The likelihood ratio test (LRT) statistic used to test sphericity was first developed by Mauchly in 1940 [1]. Although the excellent properties of the LRTs in general, it is known that the test for sphericity based on this LRT statistic may fail to detect departures from sphericity and be inaccurate when multivariate normality cannot be ensured [2, 3, 4]. These issues may be worsened when asymptotic results are required to perform the tests. The exact distribution of the sphericity LRT statistic may be represented as the distribution of the product of independent Beta random variables; however, its cumulative distribution and density functions do not have manageable expressions and thus are difficult to use in practice. For this reason, many approximations have been developed for the distribution of the LRT statistic used to test the sphericity structure. The most simple one, for the negative logarithm of the test statistic, is defined by a single chi-squared distribution and is based on Wilks theorem [5]. This approximation can be improved by multiplying the statistic by an appropriate constant [6]. These latter two approximations will be the focus of this work. As it happens with several asymptotic approximations, the chi-squared approximations are asymptotic with respect to the sample size but exhibit limitations for a large number of variables and/or when sample sizes are small. These points will be thoroughly evaluated in this study through the use of simulations.

The aim of this work is to study the properties of the sphericity test using two different approximations for the logarithm of the LRT statistic, by investigating : (1) the asymptotic behaviour of the approximations for the distributions of the test statistics under the assumption of sphericity; (2) the effect of increasing the sample size on the empirical probabilities of Type I and Type II errors of the test; and (3) the effects of dimensionality reduction on the power of the test.

This dissertation is divided in 7 chapters. In Chapter 2, we provide a theoretical background to introduce relevant concepts and ideas. Next, in Chapter 3, we derive the classic LRT statistic and introduce its improvement. Chapter 4 will focus on the simulations setup, followed by Chapter 5 in which we discuss the results. Finally, in Chapter 6, we present the conclusions that we draw from our findings and suggest future work to be carried out. Furthermore, an appendix includes the `R` code used to carry out the simulations explained in Chapter 4 and to produce the results of Chapter 5.

Chapter 2: Background Knowledge

In this chapter, we present the relevant background knowledge. In this study we consider the multivariate version of the sphericity test, thus the hypotheses are formulated based on the covariance matrix. We start by introducing the multivariate normal distribution and two alternative covariance structures. For the numerical study, we use density estimation and dimensionality reduction, hence we introduce kernel density estimation and principal components analysis.

2.1. The multivariate normal distribution

A random variable X, with observed values x, has univariate normal distribution with mean value μ and variance σ^2, i.e. we write $X \sim N(\mu, \sigma^2)$, if the probability density function (pdf) is defined by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \tag{2.1}$$

where $x, \mu \in \mathbb{R}$ and $\sigma > 0$. We can note that the term

$$\left(\frac{x-\mu}{\sigma}\right)^2 = (x-\mu)(\sigma^2)^{-1}(x-\mu) \tag{2.2}$$

represents the square of the distance between x and μ in standard deviation units and that it can be generalised for a vector $\boldsymbol{x} = (x_1, ..., x_p)$

$$(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \tag{2.3}$$

which also represents the square of a distance between $\boldsymbol{x}$ and $\boldsymbol{\mu}$.

The pdf of multivariate normal random vector with p elements, $\boldsymbol{X} = (X_1, \dots, X_p)^T$, can be simply obtained by replacing (2.2) by (2.3) and generalising the normalisation factor in (2.1)

$$\frac{1}{\sqrt{2\pi}\sigma} = (2\pi)^{-1/2}(\sigma^2)^{-1/2}$$

by a more general constant accounting for the volume density for p variables

$$\left[(2\pi)^{-1/2}\right]^p |\boldsymbol{\Sigma}|^{-1/2}.$$

Hence, a random vector $\boldsymbol{X}$ of size p has a p-dimensional normal distribution, i.e. we write $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if the pdf is defined by

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^p}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}.$$

where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ is the symmetric positive definite variance-covariance matrix [7]. That is,

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix} = \begin{pmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_p) \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{pmatrix}$$

where $\sigma_{ij} = \mathrm{Cov}(X_i, X_j)$ is the ij-th element of $\mathbf{\Sigma}$, for $i, j = 1, 2, \ldots, p$. When $i = j$, we have $\sigma_i^2 = \mathrm{Cov}(X_i, X_i) = \mathrm{Var}(X_i)$.

The multivariate normal distribution is used in many multivariate analysis techniques. Its use has several advantages, including being mathematically tractable and possessing several convenient properties, for example, (1) the distribution can be fully described using only means, variances, and covariances; (2) if the variables are not correlated, they are independent; and (3) linear functions of normal variables are also normally distributed [6].

Of course, mathematical attractiveness per se is of little use however, normal distributions are indeed useful in practice. We have two reasons for this: first, various populations are multivariate normal and even when they are not, the distribution may serve as a useful approximation; and second, the sampling distributions of many multivariate statistics are approximately normal, regardless of the form of the parent population, due to the central limit effect [8].

2.2. Covariance structure

The covariance structure provides valuable information regarding the distribution, or, in the sample case, about the data itself. In this study, we examine the properties of the sphericity test and, for the purposes of the numerical study, the compound symmetric structure is considered.

A covariance matrix $\mathbf{\Sigma}$ is said to be *compound symmetric* if all variances are equal and covariances are constant. Explicitly, for some constants $\sigma^2 > 0$ and $-\frac{1}{p-1} < \rho < 1$ we have

$$\mathbf{\Sigma} = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}. \tag{2.4}$$

A covariance matrix $\mathbf{\Sigma}$ is said to be *spherical*[1] if it is proportional to the identity matrix. A spherical matrix is a special case of a compound symmetric matrix with zero covariances. In other words, $\mathbf{\Sigma}$ is spherical if for some constant $\sigma^2 > 0$ we have

$$\mathbf{\Sigma} = \sigma^2 \boldsymbol{I} = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix}.$$

Spherical covariance matrices have $\sigma_{ij} = \mathrm{Cov}(X_i, X_j) = 0$ whenever $i \neq j$, and hence the variables $X_1, X_2, \ldots, X_p$ in $\boldsymbol{X}$ are independent[2], with constant variance σ^2 [7].

2.3. Kernel density estimation

Kernel density estimation (KDE) is the application of kernel smoothing for probability density estimation. In this study, we assess the behaviour of the asymptotic distributions of the logarithm

[1]When $\mathbf{\Sigma}$ is spherical, the ellipsoid $(\boldsymbol{X} - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) = c^2$ in the exponent of the pdf reduces to $(\boldsymbol{X} - \boldsymbol{\mu})^T (\boldsymbol{X} - \boldsymbol{\mu}) = \sigma^2 c^2$, the equation of a sphere. This is where the term *sphericity* arises from.

[2]If two variables X and Y are independent then $\mathrm{Cov}(X, Y) = 0$. However, the opposite is in general not true. In fact, $\mathrm{Cov}(X, Y) = 0$ is a necessary but not sufficient condition for the independence of X and Y.

of the test statistics. The numerical study requires the estimation of the pdf of the empirical test statistics.

Let $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ be an observed sample of a multivariate population with unknown pdf $f(\boldsymbol{x})$ at any given point $\boldsymbol{x}$ where $\{\boldsymbol{X}_i\}_{i=1}^n$ are independent and identically distributed (iid).

Then the kernel density estimator of the pdf is defined by

$$\hat{f}_h(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} K_h(\boldsymbol{x} - \boldsymbol{x}_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{\boldsymbol{x} - \boldsymbol{x}_i}{h}\right)$$

where K is the kernel (a non-negative function) and $h > 0$ is a smoothing parameter called the bandwidth [9].

In practice, the choices for K and h are handled automatically in `R` using the function `density` [10].

2.4. Principal component analysis

Principal component analysis (PCA) is a dimensionality reduction technique and an exploratory analytic tool for non-structured multivariate data. In many circumstances, the excessive number of available variables represents an additional source of unnecessary complexity difficult to deal with as, for example, when a subset of variables are highly correlated and therefore redundant. In these situations, it desirable to transform the observed variables into new and fewer variables, called principal components, loosing information as little as possible. In this study, we investigate the effects of PCA on the power of the sphericity test.

Let $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ be an observed sample of size n, where each observation $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})$ has p elements, for $i = 1, \dots, n$. We define

$$\mathbf{x}_{n\times p} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

and we wish to obtain the most interesting of all matrices within the ones with lower dimension, say $\mathbf{y}_{n\times q}$ with $q < p$.

For simplicity, let us consider a single observation of p elements $\boldsymbol{x} = (x_1, \dots, x_p)$. The sample principal components is a p-dimensional vector whose elements are the p linear combinations $\boldsymbol{y} = (y_1, \dots, y_p)$ of the p original variables $\boldsymbol{x} = (x_1, \dots, x_p)$. That is,

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p \\ y_2 &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p \\ &\vdots \\ y_p &= a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pp}x_p \end{aligned}$$

where a_{kj} are unknown coefficients to be estimated, for $k, j = 1, \dots, p$.

Define the saturations matrix (or loadings matrix) to be

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \\ \vdots \\ \boldsymbol{a}_p \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{pmatrix}$$

so that

$$y_k = a_{k1}x_1 + a_{k2}x_2 + \cdots + a_{kp}x_p = \sum_{j=1}^{p} a_{kj}x_j = \boldsymbol{a}_k^T\boldsymbol{x}$$

represents the kth principal component (PC), with mean $\bar{y}_k = \boldsymbol{a}_k^T\overline{\boldsymbol{x}}$ and variance $\mathrm{Var}(y_k) = \boldsymbol{a}_k^T\boldsymbol{S}\boldsymbol{a}_k$ where $\boldsymbol{S}$ is the sample covariance matrix, for $k = 1, \ldots, p$.

The first PC $\boldsymbol{a}_1^T\boldsymbol{x}$ is obtained by estimating the vector $\boldsymbol{a}_1$ such that it maximises the retained variance $\mathrm{Var}(\boldsymbol{a}_1^T\boldsymbol{x}) = \boldsymbol{a}_1^T\boldsymbol{S}\boldsymbol{a}_1$. It can be shown that $\boldsymbol{a}_1^T\boldsymbol{S}\boldsymbol{a}_1$ attains its maximum for $\hat{\boldsymbol{a}}_1 = \boldsymbol{e}_1$ where $\boldsymbol{e}_1$ is the first eigenvector of $\boldsymbol{S}$, with correspondent eigenvalue ℓ_1. The following PC, $\boldsymbol{a}_k^T\boldsymbol{x}$ with $k > 2$, can be determined by maximising $\boldsymbol{a}_k^T\boldsymbol{S}\boldsymbol{a}_k$ restricted to be independent of $\boldsymbol{a}_{k-1}^T\boldsymbol{x}$, which is equivalent to finding the eigen-paris $(\ell_k, \boldsymbol{e}_k)$, for $k = 1, \ldots, p$. It turns out that $\ell_k \big/ \sum_{k=1}^{p} \ell_k$ gives the proportion of the total variance explained by the k-th PC and $\sum_{k=1}^{m} \ell_k \big/ \sum_{k=1}^{p} \ell_k$ gives the proportion of the total variance explained by the first m PCs.

The process that we have just described is the repeated n times to obtain a matrix of principal components

$$\mathbf{y}_{n\times p} = \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_n \end{pmatrix} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{pmatrix}$$

and then for $q < p$ we can create an $n \times q$ matrix

$$\begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1q} \\ y_{21} & y_{22} & \cdots & y_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nq} \end{pmatrix}$$

that retains $\sum\limits_{k=1}^{q} \ell_k \Big/ \sum\limits_{k=1}^{p} \ell_k$ of the total variance [7].

Chapter 3: Likelihood Ratio Test for Sphericity

In this chapter, we derive the LRT statistic for testing sphericity. We also present the asymptotic sampling distributions considered in this study.

3.1. Likelihood ratio tests

Let $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ be a random sample of a multivariate population such that the pdf of $\boldsymbol{X}$ is $f(\boldsymbol{x}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of unknown parameters with parameter space $\boldsymbol{\Theta}$. Since $\{\boldsymbol{X}_i\}_{1=1}^n$ are iid, the joint pdf may be defined by $f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n|\boldsymbol{\theta}) = \prod_{i=1}^n f(\boldsymbol{x}_i|\boldsymbol{\theta})$ and the likelihood by $L(\boldsymbol{\theta}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n|\boldsymbol{\theta})$. Consider the space partition such that $\boldsymbol{\Theta}_0 \cup \boldsymbol{\Theta}_0^c = \boldsymbol{\Theta}$. Let $\hat{\boldsymbol{\theta}}$ be the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ and let $\hat{\boldsymbol{\theta}}_0$ be the MLE of $\boldsymbol{\theta}$ under the subspace $\boldsymbol{\Theta}_0$.

Consider the test $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ against $H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0^c$. Then under H_0, the LRT statistic is

$$\lambda(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \frac{L(\hat{\boldsymbol{\theta}}_0|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)}{L(\hat{\boldsymbol{\theta}}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)} = \frac{\sup_{\boldsymbol{\theta}_0} L(\boldsymbol{\theta}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)}{\sup_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)}. \tag{3.1}$$

The LRT has a rejection region of the form

$$R = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n | \lambda(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \leqslant c\}$$

where $c \in [0, 1]$ is an adequate constant function of the sampling distribution of $\lambda(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, under H_0, and the significance level

$$\alpha = \sup_{\boldsymbol{\theta} \in \boldsymbol{\theta}_0} \mathbb{P}_{\boldsymbol{\theta}}(\lambda(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) \leqslant c).$$

It can be shown that under strong regularity conditions, if $\{\boldsymbol{X}_i\}_{1=1}^n$ are iid, then, for sufficiently large n,

$$-2 \log \lambda(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) \overset{a}{\sim} \chi_j^2$$

where $j = r - q$, r is the number of free parameters specified by $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and q is the number of free parameters specified by $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ [5].

3.2. The Mauchly statistic

Let $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ represent an observed sample from $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The sphericity test is given by Mauchly [1] as follows:

$$H_0 : \boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_p \qquad \text{vs} \qquad H_1 : \boldsymbol{\Sigma} \neq \sigma^2 \boldsymbol{I}_p$$

with σ^2 unspecified.

Recall the definition of the LRT statistic in equation (3.1). Assuming the vector of parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we start by finding an expression for $L(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$. To simplify notation we will write $L(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) = L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. From Johnson and Wichern [7], this likelihood exists under the whole parameter space $\boldsymbol{\Theta}$ for the multivariate normal population with their joint density being

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= f(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \prod_{j=1}^{n} \left[\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}) \right\} \right] \\ &= \frac{1}{(2\pi)^{np/2}|\boldsymbol{\Sigma}|^{n/2}} \exp\left\{ -\frac{1}{2}\sum_{j=1}^{n}(\boldsymbol{x}_j - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}) \right\}. \end{aligned} \tag{3.2}$$

We now manipulate this expression to simplify the sum inside the exponent. First, from linear algebra we know that $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = \mathrm{tr}(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}) = \mathrm{tr}(\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^T)$. So we have

$$\begin{aligned} \sum_{j=1}^{n}(\boldsymbol{x}_j - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}) &= \sum_{j=1}^{n} \mathrm{tr}\Big[(\boldsymbol{x}_j - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu})\Big] \\ &= \sum_{j=1}^{n} \mathrm{tr}\Big[\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu})(\boldsymbol{x}_j - \boldsymbol{\mu})^T\Big] \\ &= \mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\sum_{j=1}^{n}(\boldsymbol{x}_j - \boldsymbol{\mu})(\boldsymbol{x}_j - \boldsymbol{\mu})^T\right]. \end{aligned}$$

We want to write the above expression in terms of the sample covariance matrix $\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T$. Note that

$$\begin{aligned} \sum_{j=1}^{n}(\boldsymbol{x}_j - \boldsymbol{\mu})(\boldsymbol{x}_j - \boldsymbol{\mu})^T &= \sum_{j=1}^{n}(\boldsymbol{x}_j - \bar{\boldsymbol{x}} + \bar{\boldsymbol{x}} - \boldsymbol{\mu})(\boldsymbol{x}_j - \bar{\boldsymbol{x}} + \bar{\boldsymbol{x}} - \boldsymbol{\mu})^T \\ &= \sum_{j=1}^{n}(\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})^T + \sum_{j=1}^{n}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^T \\ &= n\boldsymbol{S} + n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^T. \end{aligned}$$

Thus, the exponent simplifies to

$$\begin{aligned} -\frac{1}{2}\mathrm{tr}\bigg(\boldsymbol{\Sigma}^{-1}\Big[(n\boldsymbol{S}) + n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^T\Big]\bigg) &= \\ &= -\frac{1}{2}\mathrm{tr}\bigg(\boldsymbol{\Sigma}^{-1}(n\boldsymbol{S}) + n\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^T\bigg) \\ &= -\frac{1}{2}\bigg(\mathrm{tr}\Big[\boldsymbol{\Sigma}^{-1}(n\boldsymbol{S})\Big] + \mathrm{tr}\Big[n\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^T\Big]\bigg) \\ &= -\frac{1}{2}\bigg(\mathrm{tr}\Big[\boldsymbol{\Sigma}^{-1}(n\boldsymbol{S})\Big] + \Big[n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^T\Big]\bigg). \end{aligned} \tag{3.3}$$

Replacing the exponent in equation (3.2) by (3.3), we obtain the likelihood

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{np/2}|\boldsymbol{\Sigma}|^{n/2}} \exp\left\{ -\frac{1}{2}\bigg(\mathrm{tr}\Big[\boldsymbol{\Sigma}^{-1}(n\boldsymbol{S})\Big] + n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^T\bigg) \right\}. \tag{3.4}$$

The next step is to maximise this likelihood, both under H_0 and under the whole space $\boldsymbol{\Theta}$. That is, we wish to find $\sup_{H_0} L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\sup_{\boldsymbol{\mu},\boldsymbol{\Sigma}} L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, or equivalently we wish to find $L(\hat{\boldsymbol{\mu}}, \hat{\sigma^2}\boldsymbol{I}_p)$ and $L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$. The MLEs of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are well known and are defined by $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}} = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{x}_j$ and $\hat{\boldsymbol{\Sigma}} = \boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T$ [7].

So when substituting the MLEs into expression (3.4), we obtain

$$\begin{aligned}
L(\bar{\boldsymbol{x}}, \boldsymbol{S}) &= \frac{1}{(2\pi)^{np/2}|\boldsymbol{S}|^{n/2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{S}^{-1}n\boldsymbol{S})\right\} \\
&= \frac{1}{(2\pi)^{np/2}|\boldsymbol{S}|^{n/2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(n\boldsymbol{I}_p)\right\} \\
&= \frac{1}{(2\pi)^{np/2}|\boldsymbol{S}|^{n/2}} \exp\left\{-\frac{1}{2}np\right\}
\end{aligned}$$

which is precisely the supremum of the likelihood under the whole space $\boldsymbol{\Theta}$.

We now turn our attention to the case of $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}_p$, under the null hypothesis. We wish to maximise the following likelihood

$$\begin{aligned}
L(\boldsymbol{\mu}, \sigma^2\boldsymbol{I}_p) = \frac{1}{(2\pi)^{np/2}|\sigma^2\boldsymbol{I}_p|^{n/2}} \exp\Bigg\{ -\frac{1}{2}\Bigg(&\mathrm{tr}\Big[(\sigma^2\boldsymbol{I}_p)^{-1}(n\boldsymbol{S})\Big] + \\
&+ n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^T(\sigma^2\boldsymbol{I}_p)^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^T\Bigg)\Bigg\}.
\end{aligned}$$

Considering the MLE $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}$, we need to maximise

$$\begin{aligned}
L(\bar{\boldsymbol{x}}, \sigma^2\boldsymbol{I}_p) &= \frac{1}{(2\pi)^{np/2}|\sigma^2\boldsymbol{I}_p|^{n/2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\Big[(\sigma^2\boldsymbol{I}_p)^{-1}(n\boldsymbol{S})\Big]\right\} \\
&= \frac{1}{(2\pi)^{np/2}(\sigma^2)^{np/2}} \exp\left\{-\frac{n}{2\sigma^2}\mathrm{tr}(\boldsymbol{S})\right\}
\end{aligned}$$

with respect to $\sigma^2\boldsymbol{I}_p$, which equates to maximising with respect to σ^2. We proceed by using the standard technique to find the MLE for σ^2 by differentiating the log-likelihood with respect to σ^2 and then setting the derivative to zero:

$$\begin{aligned}
\log L(\bar{\boldsymbol{x}}, \sigma^2\boldsymbol{I}_p) &= \log\left(\frac{1}{(2\pi)^{np/2}(\sigma^2)^{np/2}} \exp\left\{-\frac{n}{2\sigma^2}\mathrm{tr}(\boldsymbol{S})\right\}\right) \\
&= -\log\Big[(2\pi)^{np/2}(\sigma^2)^{np/2}\Big] - \frac{n}{2\sigma^2}\mathrm{tr}(\boldsymbol{S}) \\
&= -\frac{np}{2}\log(2\pi) - \frac{np}{2}\log(\sigma^2) - \frac{n}{2\sigma^2}\mathrm{tr}(\boldsymbol{S})
\end{aligned}$$

and then by differentiating and setting equal to zero,

$$\begin{aligned}
&\left.\frac{\partial \log L(\bar{\boldsymbol{x}}, \sigma^2\boldsymbol{I}_p)}{\partial\sigma^2}\right|_{\hat{\sigma^2}} = 0 \\
\implies &\left. -\frac{np}{2\sigma^2} + \frac{n}{2(\sigma^2)^2}\mathrm{tr}(\boldsymbol{S})\right|_{\hat{\sigma^2}} = 0 \\
\implies &-\frac{p}{\hat{\sigma^2}} + \frac{\mathrm{tr}(\boldsymbol{S})}{(\hat{\sigma^2})^2} = 0 \\
\implies &\frac{p}{\hat{\sigma^2}} = \frac{\mathrm{tr}(\boldsymbol{S})}{(\hat{\sigma^2})^2} \\
\implies &\hat{\sigma^2} = \frac{\mathrm{tr}(\boldsymbol{S})}{p}.
\end{aligned}$$

It can be proved that this is indeed the maximum [7] by showing that the second derivative is positive. We are now finally able to calculate the supremum of the likelihood under H_0:

$$\begin{aligned}
L(\bar{\boldsymbol{x}}, \hat{\sigma^2}\boldsymbol{I}_p) &= \frac{1}{(2\pi)^{np/2}(\hat{\sigma^2})^{np/2}} \exp\left\{-\frac{n}{2\hat{\sigma^2}}\mathrm{tr}(\boldsymbol{S})\right\} \\
&= \frac{1}{(2\pi)^{np/2}(\mathrm{tr}(\boldsymbol{S})/p)^{np/2}} \exp\left\{-\frac{n}{2(\mathrm{tr}(\boldsymbol{S})/p)}\mathrm{tr}(\boldsymbol{S})\right\} \\
&= \frac{1}{(2\pi)^{np/2}(\mathrm{tr}(\boldsymbol{S})/p)^{np/2}} \exp\left\{-\frac{1}{2}np\right\}
\end{aligned}$$

and so the LRT statistic for the spherical test for the multivariate normal distribution is

$$\lambda(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) = \frac{L(\bar{\boldsymbol{x}}, \hat{\sigma^2}\boldsymbol{I}_p)}{L(\bar{\boldsymbol{x}}, \boldsymbol{S})} = \frac{\frac{1}{(2\pi)^{np/2}(\mathrm{tr}(\boldsymbol{S})/p)^{np/2}} \exp\left\{-\frac{1}{2}np\right\}}{\frac{1}{(2\pi)^{np/2}|\boldsymbol{S}|^{n/2}} \exp\left\{-\frac{1}{2}np\right\}} = \left[\frac{|\boldsymbol{S}|}{(\mathrm{tr}(\boldsymbol{S})/p)^p}\right]^{n/2}.$$

which is the statistic Mauchly derived in 1940 [1].

3.3. Asymptotic sampling distributions

As mentioned, it can be shown that, under strong regularity conditions, if $(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)$ are iid vectors under H_0, then, for large n,

$$-2\log\lambda(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) \overset{a}{\sim} \chi^2_j$$

where j is the total number of parameters minus the number estimated by the restrictions imposed by the null hypothesis [5].

Let $U = -2\log\lambda(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$, then

$$\begin{aligned}
U &= -2\log\lambda(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) \\
&= -2\log\left[\frac{|\boldsymbol{S}|}{(\mathrm{tr}(\boldsymbol{S})/p)^p}\right]^{n/2} \\
&= -2\cdot\frac{n}{2}\log\left[\frac{|\boldsymbol{S}|}{(\mathrm{tr}(\boldsymbol{S})/p)^p}\right] \\
&= -n\log U^*
\end{aligned}$$

with $U^* = \dfrac{|\boldsymbol{S}|}{(\mathrm{tr}(\boldsymbol{S})/p)^p}$. Then,

$$U = -n\log U^* \overset{a}{\sim} \chi^2_j$$

with $j = \frac{1}{2}p(p+1) - 1$, since under the whole space, we need to compute $\boldsymbol{S}$ and the total number of parameters to be estimated is $p + \binom{p}{2} = \frac{1}{2}p(p+1)$, and under the null hypothesis we have the restriction $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}_p$, so we only need to estimate σ^2.

Rencher and Christensen [6] consider a slight variation of the U test statistic which allegedly improves the asymptotic convergence to the chi-squared distribution and that is defined as

$$U' = -\left(n - 1 - \frac{2p^2 + p + 2}{6p}\right)\log U^* \overset{a}{\sim} \chi^2_j$$

with $j = \frac{1}{2}p(p+1) - 1$.

Chapter 4: Simulations

In this chapter, we will describe the simulation setup to address the following challenges:

Challenge 1 Study the asymptotic approximations for the distributions of U and U' under the assumption of sphericity.

Challenge 2 Evaluate empirically the effect of increasing sample size on the probabilities of Type I and Type II errors of the sphericity test considering U and U'.

Challenge 3 Investigate the effects of dimensionality reduction on the power of the sphericity test considering U and U'.

In the first challenge, we want to study how the empirical distributions of U and U' change as the sample size n increases. Under H_0, we assume a spherical multivariate normal population $N_p(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_p)$. For simplicity we consider $\boldsymbol{\mu} = \mathbf{0}$ and $\sigma^2 = 1$ and we choose $p = 10$ as it is large enough to show the effects of dimensionality reduction in challenge 3 but not too large that it requires an unfeasible amount of computational power. We choose n ranging from 11 to 21 by a step size of 2. We start at 11, as we need $n > p$, to make sure that $\log u$ is well-defined, we end at 21 as values beyond that show slower convergence to the asymptotic approximation, and we have a step size of 2 to save computational time and declutter plots.

To obtain the empirical distributions of U and U', we take n random observations from a normal population $N_{10}(\mathbf{0}, \boldsymbol{I}_{10})$, i.e. we generate a random sample of size n from $N_{10}(\mathbf{0}, \boldsymbol{I}_{10})$, and we replicate the process 10000 times (Algorithm 1, line 3). We choose 10000 repetitions as a good trade-off between accuracy of results and computational time. Within each replication, and for each sample of size n, we calculate the values of U and U' and store them (Algorithm 1, line 4). Then the collection of observed values for U and U' allow to define their empirical distributions and KDE is used to estimate their densities (Algorithm 1, line 6). These densities are then plot against their asymptotic distribution χ^2_{54} distribution[1].
To quantify how different the estimated densities are to the asymptotic χ^2_{54} distribution, we define a 'distance' measure as the average of the absolute value of the difference between the quantiles of the χ^2_{54} and the empirical distributions. We repeat this process for different values of n. The obtained results and plots are shown in Section 5.1.

To investigate the long term behaviour of the 'distance' between the empirical and the theoretical asymptotic distributions, we used a different simulation set-up, letting n range from 11 to 81 with a step size of 10. Algorithm 1 summarises how we obtain the empirical distributions. The analogue `R` code used in the simulations can be found in Appendix A.1.

In challenge 2, we are interested in the behaviour of the probabilities of Type I and Type II errors of the sphericity test, as n increases. A Type I error occurs when we reject the null hypothesis given it is true and a Type II error occurs when we fail to reject the null hypothesis given it is false. Let the probabilities of a Type I and Type II error be α and β, respectively. We further define the power of a test as the probability of rejecting the null hypothesis when it is false, i.e Power $= 1 - \mathbb{P}(\text{Type II error}) = 1 - \beta$.

To estimate α, we once again generate a random sample of size n from $N_{10}(\mathbf{0}, \boldsymbol{I}_{10})$ and replicate the process 10000 times (Algorithm 2, line 4). Using the U and U' statistics, for each sample

[1] The degrees of freedom are given by $p(p+1)/2 - 1 = 10 \times 11/2 - 1 = 54$.

Algorithm 1 Computation of the empirical distributions of U and U' varying sample size n

```
for n from 11 to 21 by 2 do
    for 10000 times do
        draw a random sample of size n from N_10(0, I_10)
        compute the observed values for U and U' (U_obs and U'_obs, respectively)
    end for
    define the empirical distributions of U and U' for sample size n
end for
```

of size n, we carry out the sphericity test at the α significance level (Algorithm 2, line 5). Our simulation uses the 0.05 significance level. If the observed value of the U statistic exceeds the $(1-\alpha)$th chi-square quantile, i.e., $U_{obs} > \chi^2_{54;(1-\alpha)}$ or similarly $U'_{obs} > \chi^2_{54;(1-\alpha)}$, then we reject H_0 (Algorithm 2, line 6). We then estimate the α value as the proportion of times we rejected H_0 out of 10000 (Algorithm 2, line 9).

The power is estimated in a similar way by generating 10000 random samples of size n from a multivariate normal population under $H_1 : \boldsymbol{\Sigma} \neq \sigma^2 \boldsymbol{I}$. There are many ways we could consider $N_{10}(\mathbf{0}, \boldsymbol{\Sigma})$ such that $\boldsymbol{\Sigma} \neq \sigma^2 \boldsymbol{I}$. We choose to work with compound-symmetric covariance matrices $\boldsymbol{\Sigma}$ as defined in (2.4) for $\rho \in [-1, 1]$. Since the null hypothesis occurs at $\rho = 0$, we can quantify the departure from H_0 as $|\rho - 0|$. We are then able to estimate the power and plot it against ρ, which highlights its departure from H_0. Note that the estimated power of the test at $\rho = 0$ is expected to correspond to the α value.

To calculate the power, we generate 10000 random samples of size n from a $N_{10}(\mathbf{0}, \boldsymbol{\Sigma}_\rho)$ population, with $\rho \in [-0.1, 0.25]$ as the estimated power isn't expected to change considerably outside these values and computational time constraints limit the step size to 0.0175[2] (Algorithm 2, line 4). As before, we carry out the sphericity test for each sample and the power is estimated as the proportion of times we reject H_0 out of 10000 (Algorithm 2, lines 5, 6, and 8). The power curve is the resulting plot of the estimated power against ρ (Algorithm 2, line 11).

We repeat this process for different values of n (this time from $n = 11$ to $n = 81$ with a step size of 10, to better show convergence) and the results and plots are shown in Section 5.2. Algorithm 2 summarises how we can estimate α and $1 - \beta$. The analogue `R` code used in the simulations can be found in Appendix A.2.

Algorithm 2 Computation of the power of the shericity test varying sample size n

```
1:  for n from 11 to 81 by 10 do
2:      for ρ from -0.1 to 0.25 by 0.0175 do
3:          for 10000 times do
4:              draw a random sample of size n from N_10(0, Σ_ρ)
5:              perform the sphericity test using U_obs and U'_obs
6:              store whether we rejected H_0 or not
7:          end for
8:          compute the power of the test based on U and U', for ρ and sample size n
9:          compute the alpha of the test based on U and U', only if ρ = 0, and sample size n
10:     end for
11:     plot the estimated power against ρ for sample size n
12: end for
```

[2]The step size is obtained by dividing the interval into 20 evenly spaced parts, this is given by $[0.25 - (-0.1)]/20 = 0.0175$.

Finally, for the last challenge, we want to investigate the effects of dimensionality reduction on the power of the sphericity test. We start by calculating the power in the same way as before. We generate 10000 random samples of size n from a $N_{10}(\mathbf{0}, \mathbf{\Sigma}_\rho)$ population, with ρ ranging from 0 to 0.25 by a step size of 0.0175 (Algorithm 3, line 4). We add the constraint $\rho > 0$ to avoid numerical issues with dimensionality reduction. Additionally, for each sample we perform PCA and retain $10 - j$ PCs, for $j = 1, 2, 3$ (Algorithm 3, line 8), and then calculate the power (Algorithm 3, line 13). We obtain a power curve for the statistics considering the retained $10 - j$ PCs, for $j = 1, 2, 3$ (Algorithm 3, line 15).

We repeat this process for different values of n (this time from $n = 11$ to $n = 91$ with a step size of 20, to better show the effects of dimensionality reduction) and the results and plots will be shown in Section 5.3. Algorithm 3 summarises how we can obtain the power curves. The analogue `R` code used in the simulations can be found in Appendix 3.

Algorithm 3 Computation of the power of the sphericity test when performing PCA varying sample size n

1: **for** n from 11 to 81 by 10 **do**
2:     **for** ρ from 0 to 0.25 by 0.0175 **do**
3:         **for** 10000 times **do**
4:             generate a random sample of size n from $N_{10}(\mathbf{0}, \mathbf{\Sigma}_\rho)$
5:             carry out a sphericity test using U_{obs} and U'_{obs}
6:             store whether we rejected H_0 or not
7:             **for** $j = 1, 2, 3$ variables removed **do**
8:                 perform PCA and retain $10 - j$ PCs
9:                 carry out a sphericity test using U_{obs} and U'_{obs}
10:                 store whether we rejected H_0 or not
11:             **end for**
12:         **end for**
13:         calculate the power of U and U' considering the retained $10 - j$ PCs, for $j = 0, 1, 2, 3$ (for this specific ρ and n)
14:     **end for**
15:     plot the estimated power against ρ (for $j = 0, 1, 2, 3$ and this specific n)
16: **end for**

Chapter 5: Results

In this chapter we present the results from the simulations that we described in the previous chapter. The chapter is organized into 3 sections summarizing respectively the obtained results to answerer challenges 1 (section 5.1), 2 (section 5.2), and 3 (section 5.3)

5.1. Asymptotic sampling distributions under the assumption of sphericity

Figures 5.1 and 5.2 show, respectively, the asymptotic distributions of U and U' and the estimated densities across different sample sizes n. The black line represents the density of the χ^2_{54} distribution, and the coloured curves show the empirical density estimates of the statistics over the previously specified sample sizes. In both figures we observe that as n increases, the densities shift leftwards and tend to approximate the χ^2_{54} distribution. However, there are some differences. The distributions of U tend to be more right-skewed and deviate significantly from the χ^2_{54} distribution, especially for small n. On the other hand, the distributions of U' converge quicker to the χ^2_{54} distribution and the deviation from its asymptotic approximation is not as extreme as in the case for the U statistic, even with for small values of n.

Figure 5.3 displays a grid of histograms of U_{obs} and U'_{obs} across different sample sizes n superimposed on the χ^2_{54} distribution. In the right column we see that by $n = 17$, the empirical distribution of the U' statistic is well approximated by the χ^2_{54} distribution while, on the left column, the empirical distribution of U is not well approximated even for $n = 21$. For small n, we also observe that both histograms for U and U' deviate substantially from their asymptotic approximation, noting that U' offers a slightly better approximation.

The plot in Figure 5.4 depicts the change of the distance (as explained in Chapter 4) of the estimated densities of U (in red) and U' (in black) to their χ^2_{54} approximation, with sample size. For both statistics the distances decrease as n increases. The distance of U' is always smaller than the distance of U and for small n the distance for U is much bigger than the distance for U'. Figure 5.5 shows a similar plot, with n now ranging from 11 to 81. What we observed before remains true and, additionally, one can see that the gap between the distances for U and U' decrease, as n increases.

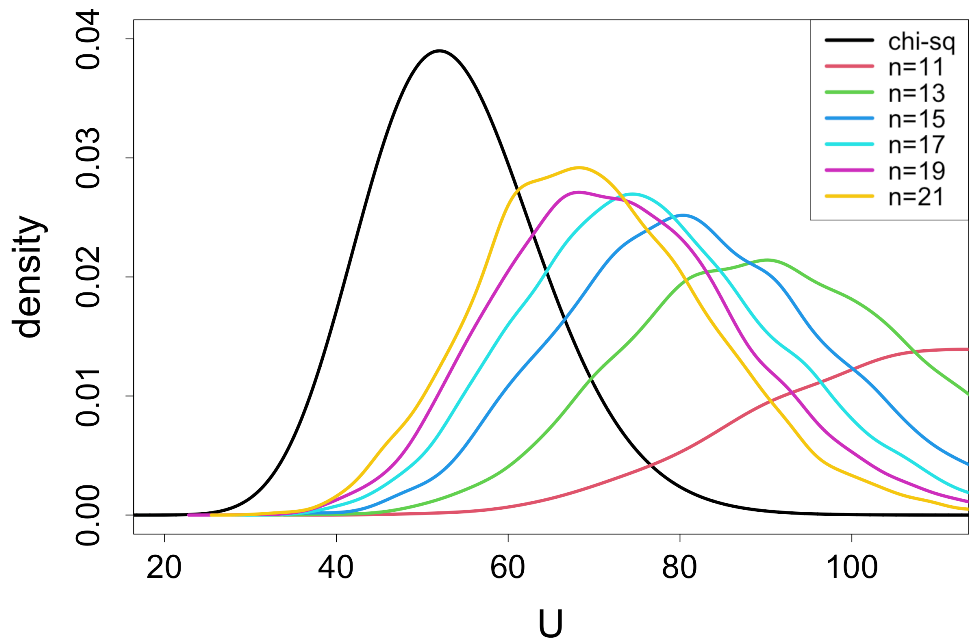

Figure 5.1: Asymptotic distribution (χ^2_{54}, black line) and estimated densities of the U statistic over various sample sizes n.

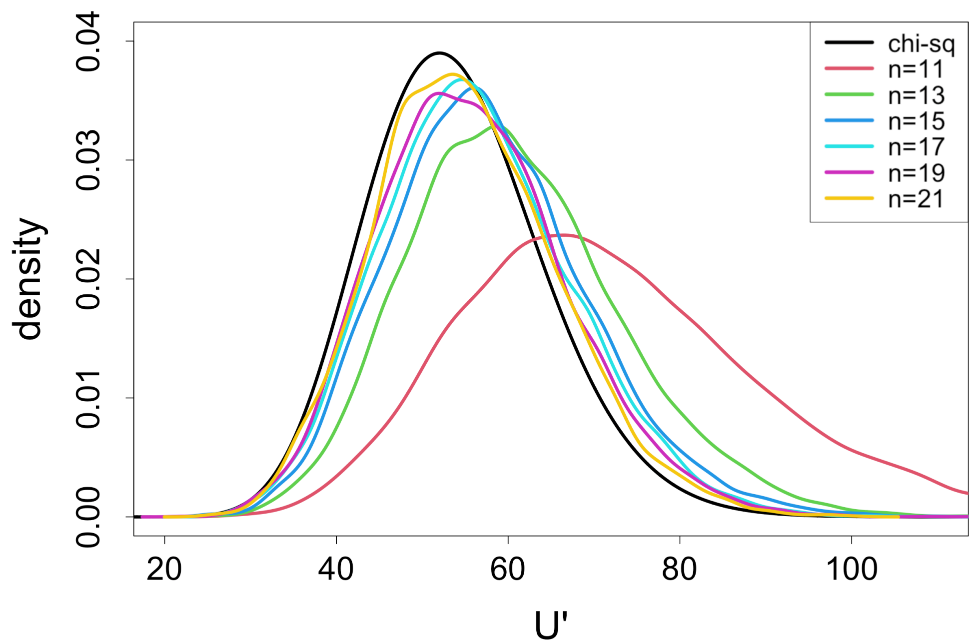

Figure 5.2: Asymptotic distribution (χ^2_{54}, black line) and estimated densities of the U' statistic over various sample sizes n.

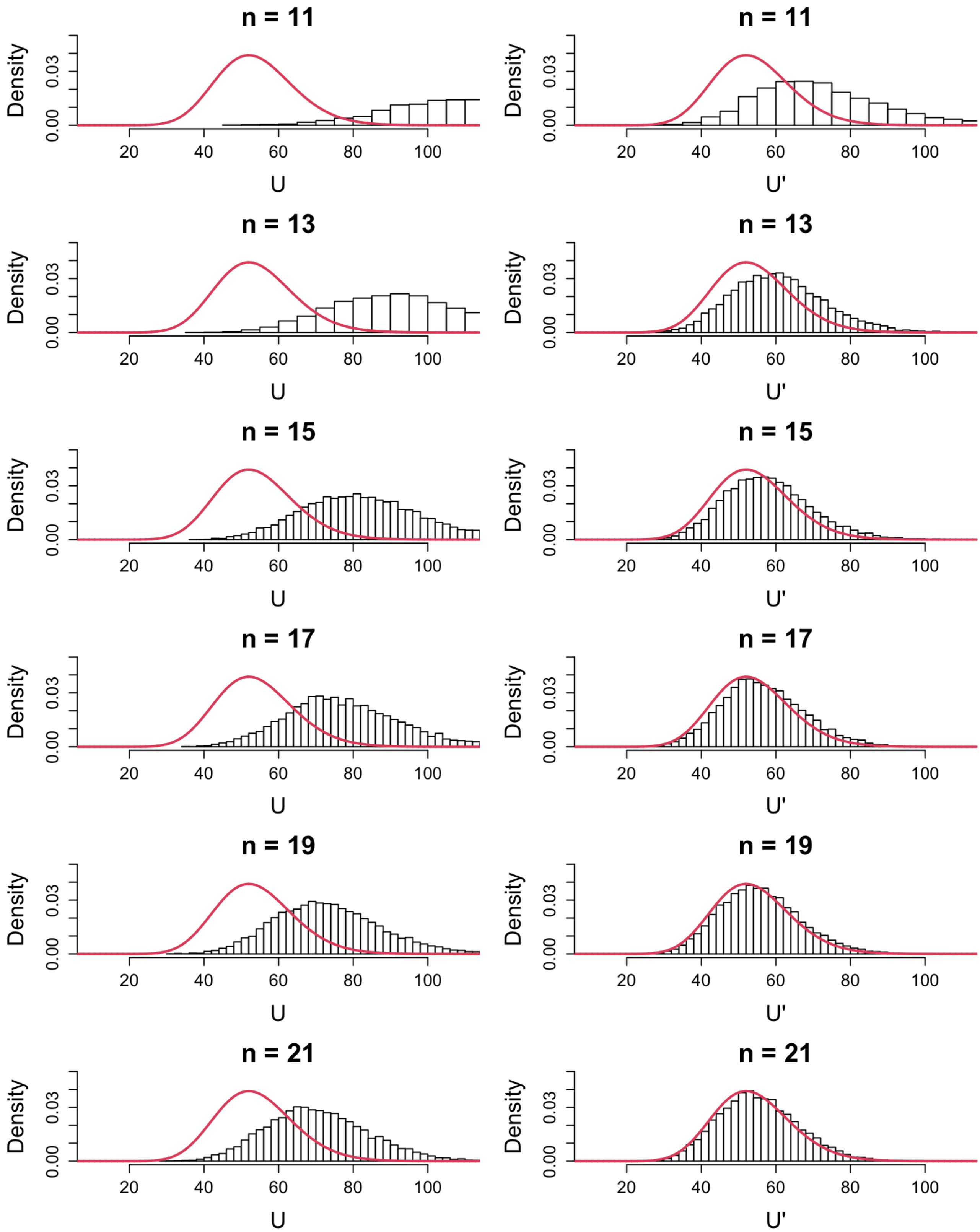

Figure 5.3: Grid of histograms of the empirical distributions of U and U' superimposed by the asymptotic distribution (χ^2_{54}, red line) across different sample sizes n.

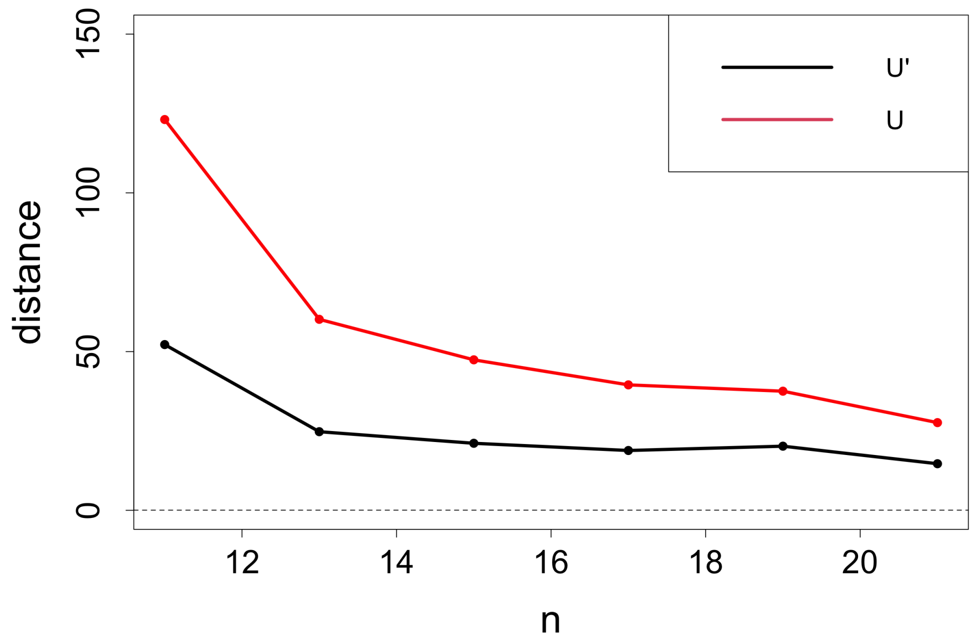

Figure 5.4: Distance between the estimated densities of the U and U' statistics to the density of the asymptotic distribution (χ^2_{54}) against the sample size n, ranging from $n = 11$ to $n = 21$.

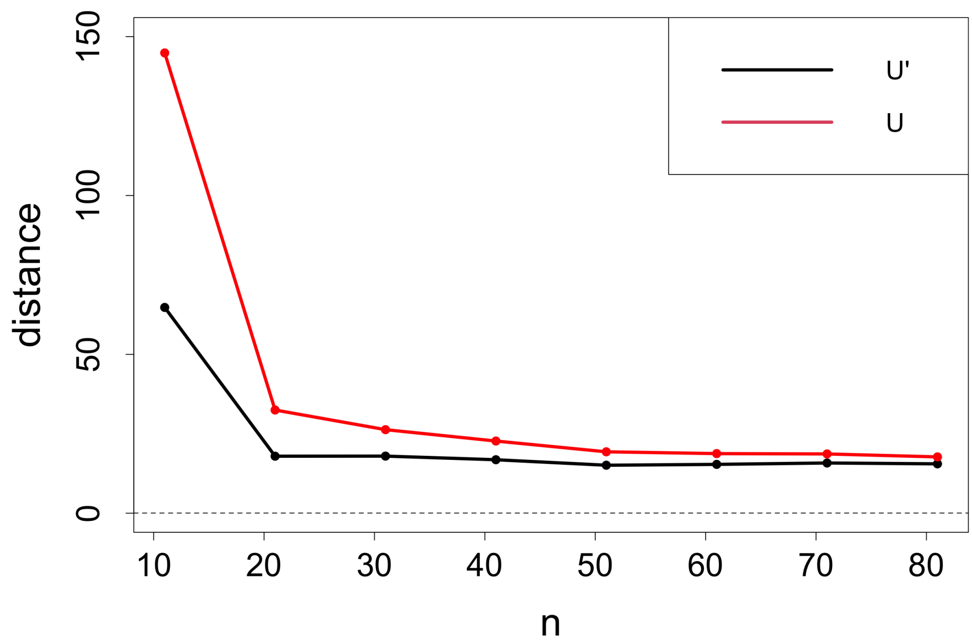

Figure 5.5: Distance between the estimated densities of the U and U' statistics to the density of the asymptotic distribution (χ^2_{54}) against the sample size n, ranging from $n = 11$ to $n = 81$.

5.2. Sample size effect on the probabilities of Type I and Type II errors of the sphericity test

Figure 5.6 shows how the estimated α values of the sphericity test using U (in red) and U' (in black) vary across different values of n. For both statistics, the estimated α values decrease as n increases and tend to approximate the significance level of the test (indicated by a dashed line). The $\hat{\alpha}$ values when considering U converge slower to α than when considering U'. For low n, $\hat{\alpha}$ is quite high using both statistics, but it is particularly high for the statistic U.

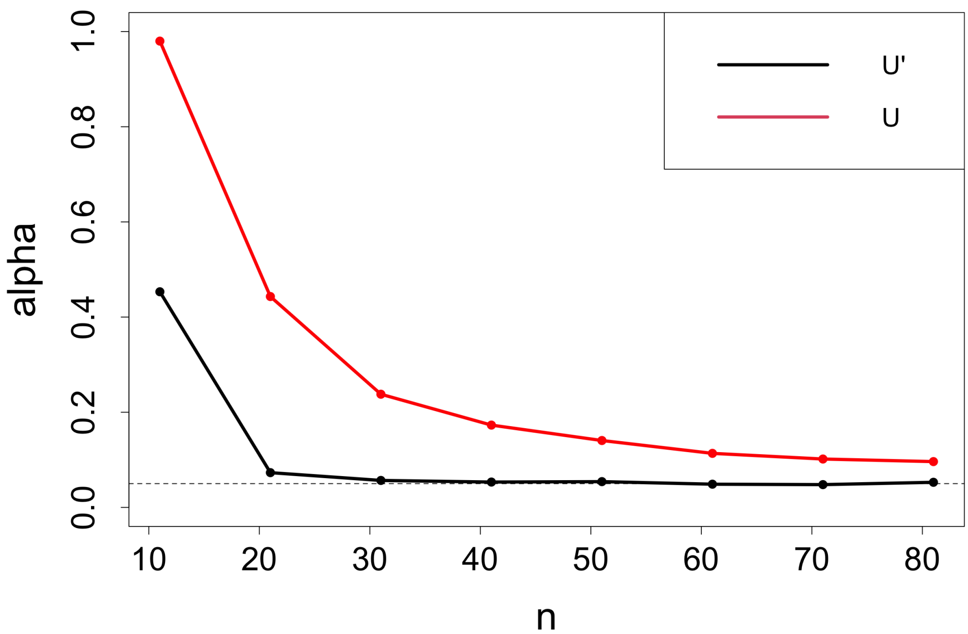

Figure 5.6: Estimated α values of the sphericity test using the U and U' statistics against the sample size n.

Figures 5.7 and 5.8 show the estimated power curves of the sphericity test using the U and U' statistics, respectively, for different n. Both figures show that the lowest estimated power occurs near $\rho = 0$, then the power increases as $|\rho|$ increases and the power increases more sharply for negative ρ. For the statistic U, the estimated power of the test decreases near $\rho = 0$, as n increases. Furthermore, at $\rho = 0$, the estimated power is equivalent to the estimated α value and it doesn't converge to the significance level (indicated by a dashed line). However, for the U' statistic, excluding very low n, the estimated power increases everywhere as n increases and at $\rho = 0$, the alpha values converges to the significance level of the test.

Figure 5.9 is a grid of plots comparing the estimated power curves of the sphericity test using U and U' across different sample sizes. Both power curves show that the lowest estimated power occurs near $\rho = 0$ and the power increases quicker for $\rho > 0$ than for $\rho < 0$. The estimated power using U is always greater than using U', but for $\rho = 0$ this means that the estimated α value of the test using U is always greater than using U'. We also observe that already by $n = 21$ at $\rho = 0$, the estimated power curve using U' reaches the significance level of the test, while using U we don't reach the significance level of the test.

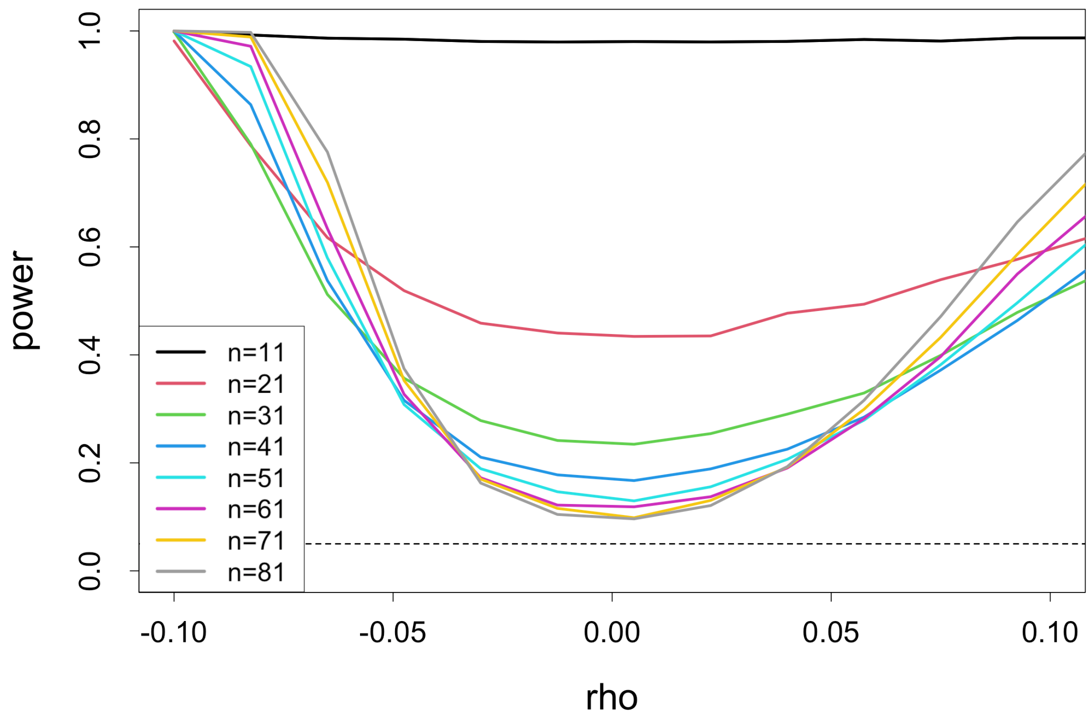

Figure 5.7: Estimated power curves of the sphericity test using the U statistic for various sample size n.

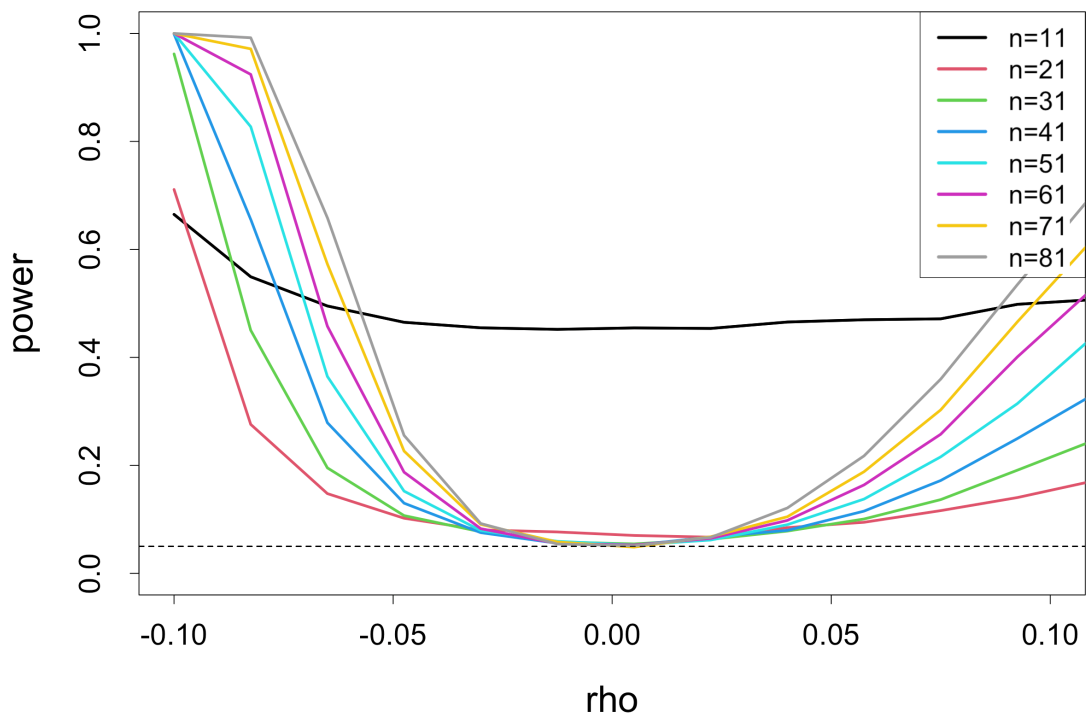

Figure 5.8: Estimated power curves of the sphericity test using the U' statistic for various sample size n.

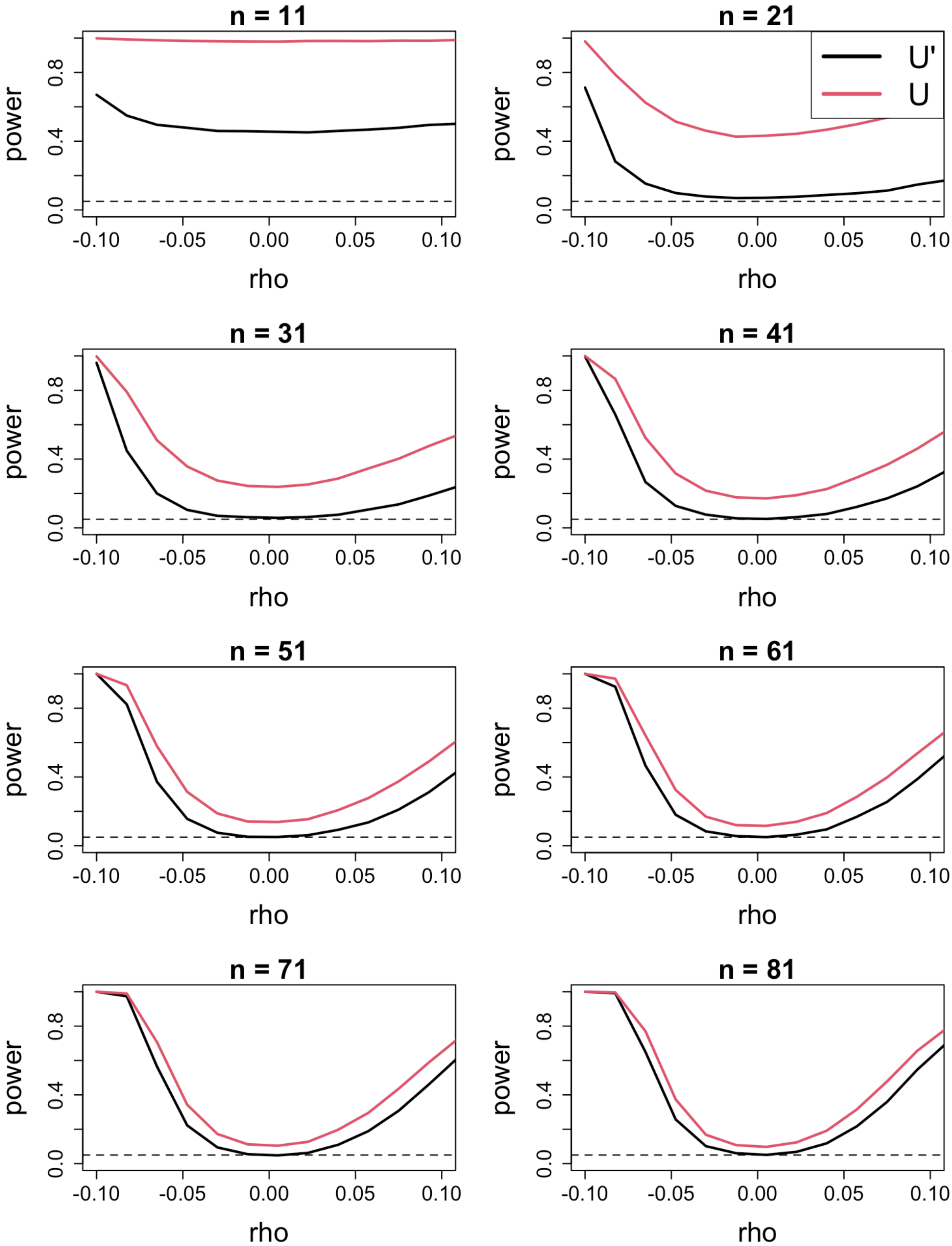

Figure 5.9: Grid of plots of the estimated power curves of the sphericity test using the statistics U and U' across different sample sizes n.

5.3. Dimensionality reduction effect on the power of the sphericity test

Figure 5.10 presents a matrix of plots of estimated power curves for the sphericity test using U and U', when retaining 10, 9, 8, and 7 PCs across different sample sizes n. For each plot, we observe that there is an increasing loss of power by retaining less PCs. As the sample size increases, the power loss due to dimensionality reduction decreases. For low values of n, using the U' statistic leads to a better retention of power than using U. However, this trend is not clear for larger n.

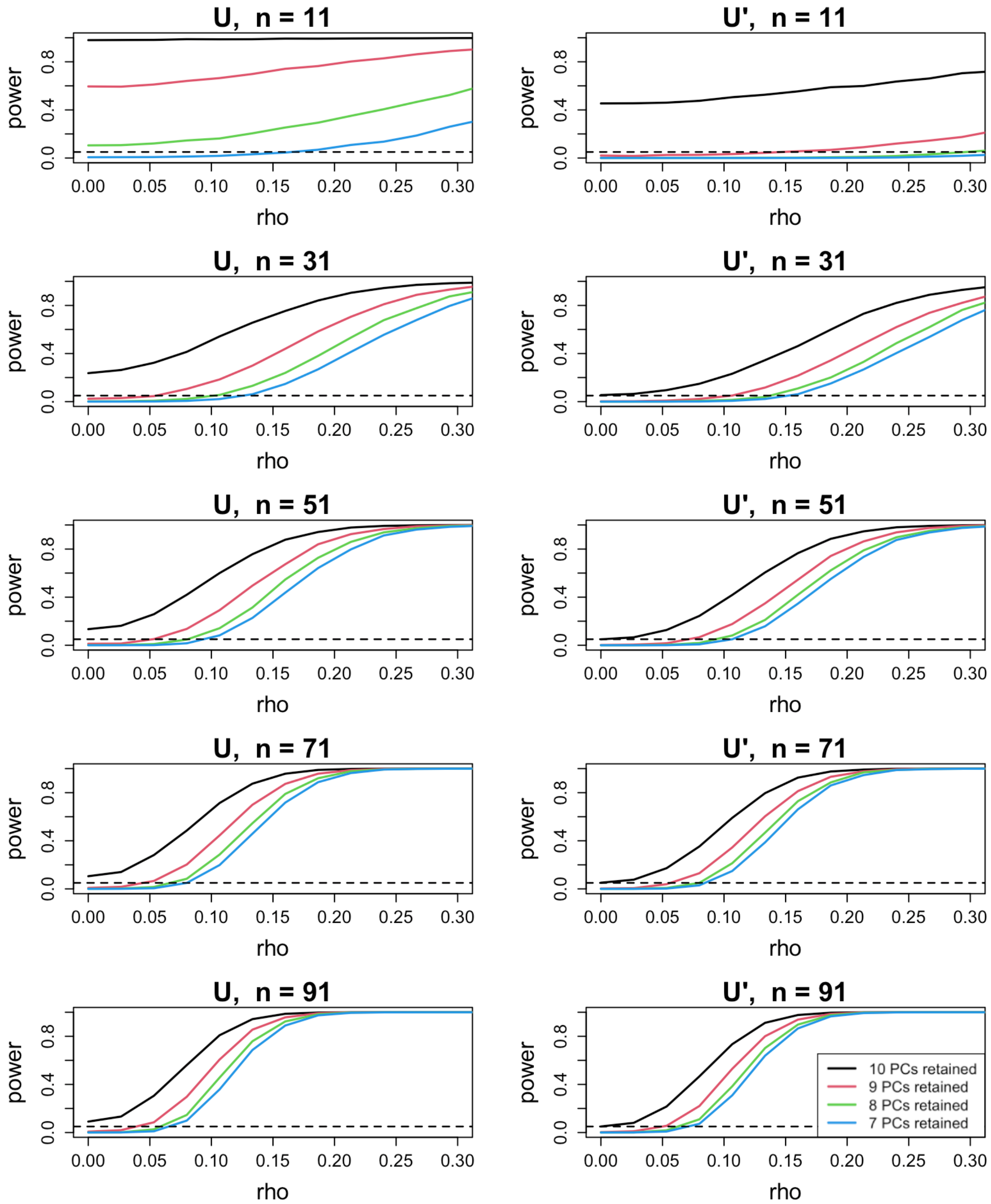

Figure 5.10: Estimated power curves of the sphericity test using the statistics U and U' over a dimension reduced space retaining 10, 9, 8, and 7 PCs, across different sample sizes n.

Chapter 6: Conclusion

In this study, we investigate the properties of the sphericity test considering the statistics U and U'. To compare their behaviour, we use three criteria, corresponding to the three challenges described in Section 4. We next highlight the main conclusions that emerge from the described study.

Regarding the asymptotic behaviour of both statistics, we conclude that their empirical distributions clearly fail to provide a good approximation of the chi-squared distribution, if the sample size is close to the number of variables. However, as the sample size increases to around 20, the estimated distribution of U' already approximates very well the theoretical distribution, while the estimated distribution of U still fails to provide a good approximation. Our simulations also show that the quality of the approximations becomes similar for large sample sizes (i.e., n greater than 80). Thus, this first criterion points to recommend the use of the statistic U', even if it provides only a slight improvement over U, since the increase in complexity of U' is negligible.

From the above, it may seem that the use of U' only has a clear advantage over the use of U if the sample size is small. However, when considering the second criterion, based on the respective power curves, the difference between the two statistics becomes more apparent. Our results show that for small sample sizes there is an overestimation of the power of the test which indicates an over-rejection process, regardless the statistic considered. It is then plausible to infer that the use of small sample sizes doesn't provide enough information to allow for a correct decision and, at the same time, compromises the precision of the estimated quantiles. Increasing the sample size shows two combined effects: it directly increases the amount of information available to make a correct inference about the true covariance structure, and it indirectly improves the approximation for the quantiles, which is in accordance to the first criterion. We also can conclude that regardless of the sample size, the estimated power using the U statistic is always greater than the one using U', showing a consistent over-rejection process. Furthermore, the power of the test fails to achieve the significance level when using the statistic U, even for large sample sizes. In contrast, the use of U' statistic allows to achieve the true significance level. Thus, this second criterion again suggests the use of the statistic U' as a general recommendation.

The conclusions highlighted until now clearly favour the use of U' to test sphericity. However, in our last study, the two statistics behave similarly, especially for large sample sizes. We find that when performing dimensionality reduction, even if we remove a small subset of variables, i.e. $p \leqslant 3$, the power of the test based on both statistics decreases substantially. It should be noted, however, that this behaviour improves by increasing the sample size. In fact, when the sample size is greater than 90, and for positive correlations above 0.2, the estimated power of the test holds even when reducing the number of retained PCs. We note that for low positive correlation of $\rho < 0.2$, the retained power decreases as more PCs are removed. The effect on power when using U or U' is similar and we observed results that we didn't expect, namely a value of estimated power for small ρ that is lower than the significance level of the test.

In summary, when carrying out a sphericity test under the assumption of a multivariate normal parent population, we suggest the use of the statistic U', instead of U, since (1) its distribution shows a better approximation to the expected asymptotic behaviour, (2) it allows to, given a fixed significance level, achieve a lower probability of incurring in a Type II error, and (3) the correspondent increase in complexity is negligible.

We end this chapter highlighting some interesting questions, still to be answered, that emerged from the studies we carried out. In our simulations, we fixed the number of variables p while varying the sample size n. Future work could include investigating the relationship between n and p to find out what ratio achieves a fair approximation for the statistics' distribution. It could also be interesting to compute the rates of convergence of the asymptotic distribution. In the second challenge, we observed asymmetric power curves depending on the sign of the correlation. Future work may include exploring the reasons why this is the case, for example if it depends on the value of σ^2 assumed in the simulations to define the covariance matrix. Additionally, it would also be interesting to study the effect of other covariance structures. Finally, the last challenge provides many open questions. For example, it could be worth looking into how using different dimensionality reduction methods (such as, e.g., UMAP[1], t-SNE[2], NMF[3] [9]) would affect the results. We conclude that more work needs to be done to understand the behaviour of the estimated power of the sphericity test and whether sphericity is preserved when dimensionality reduction is performed.

[1]Uniform manifold approximation and projection

[2]T-distributed Stochastic Neighbor Embedding

[3]Non-negative matrix factorization

Bibliography

[1] John W. Mauchly. Significance Test for Sphericity of a Normal n-Variate Distribution. *The Annals of Mathematical Statistics*, 11(2):204–209, 1940.

[2] Robert J. Boik. A Priori Tests in Repeated Measures Designs: Effects of Nonsphericity. *Psychometrika*, 46(3):241–255, 1981. Place: Chicago : Publisher: Psychometric Society,.

[3] H. J. Keselman, Joanne C. Rogan, Lawrence J. Breen, and Jorge L. Mendoza. Testing the Validity Conditions of Repeated Measures F Tests. *Psychological Bulleting*, 87(3):479–481, 1980.

[4] John E. Cornell, Dean M. Young, Samuel L. Seaman, and Roger E. Kirk. Power Comparisons of Eight Tests for Sphericity in Repeated Measures Designs. *Journal of Educational Statistics*, 17(3):233–249, 1992.

[5] S. S. Wilks. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

[6] Alvin C. Rencher and William F. Christensen. *Methods of Multivariate Analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2012.

[7] Richard A. Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall, 2007.

[8] David J. Olive. *Statistical Theory and Inference*. Springer International Publishing, 2014.

[9] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer International Publishing, 2009.

[10] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

Appendix A: R Code

We will require the following packages.

```
library(MASS)
library(mixtools)
library(matrixcalc)
library(car)
inc <- function(x) {
 eval.parent(substitute(x <- x + 1))
}
```

Below is the code to calculate the test statistics U and U'.

```
u0 <- function(data) {
  # sample covariance matrix
  S <- cov(data)
  # number of variables
  p <- ncol(data)
  # number of observations
  n <- nrow(data)

  u <-p^p * det(S) / (sum(diag(S)))^p
  return(-n * log(u))
}

u1 <- function(data) {
# sample covariance matrix
  S <- cov(data)
# number of variables
  p <- ncol(data)
# u and statistic u'
  u <-p^p * det(S) / (sum(diag(S)))^p
# number of observations
  n <- nrow(data)

  u1 <- -(n-1 - ((2* p^2 +p +2)/( 6* p))) * log(u)
  return(u1)
}
```

A.1. R Code for Challenge 1

The code used to run the simulation is the following.

```
u0_distr_n <- list()
u0_alpha_n <- list()

u1_distr_n <- list()
u1_alpha_n <- list()
k = 1
```

```
num_of_obs <- seq(from=11,to=21,by=2)


for(N in num_of_obs) {
  nboot <- 10000 # Number of bootstraps
  alpha <- 0.05 # Significance level
  P <- 10 # Number of variables
  degfreed <- 1/2 * P * (P+1) -1 #degrees of freedom

  # Parameters for multivariate normal distribution, under H0.
  mu <- rep(0,P) # Mean vector
  sigma <- diag(P) # Covariance matrix

  u0_distr <- rep(0,nboot)
  u1_distr <- rep(0,nboot)

  for (i in 1:nboot) {
    mvn <- mvrnorm(N, mu = mu, Sigma = sigma ) # from MASS package
    u0_distr[i] <- u0(mvn)
    u1_distr[i] <- u1(mvn)

  }

  u0_distr_n[[k]] <-  u0_distr
  u0_alpha_n[[k]] <- (length(which(u0_distr_n[[k]] >
                      qchisq(1-alpha,degfreed)))) / length(u0_distr_n[[k]])
  u1_distr_n[[k]] <-  u1_distr
  u1_alpha_n[[k]] <- (length(which(u1_distr_n[[k]] >
                      qchisq(1-alpha,degfreed)))) / length(u1_distr_n[[k]])
  inc(k)
}
```

Below is the code for the plots.

```
##### Plots for u0 (swap u0 with u1 to obtain the plots for u1)#####
index = 2

# Defines the parameters of the plot
x.grid <- seq(0,120,length = 10000)
chisqcurve <- dchisq(x.grid, df = degfreed)


# Histogram of the distribution of u0 for one specific population size,
# overlapped by theoretical curve
par(pty="s")
hist(u0_distr_n[[index]], prob = TRUE, col = "white",
     ylim = c(0, max(chisqcurve)), breaks = 50)
lines(x.grid, chisqcurve, col = 2, lwd = 2)
# Produces a Q-Q plot
theoretical_u0 <- qchisq(ppoints(length(u0_distr_n[[index]])),
                        df = degfreed)
qqplot(theoretical_u0,u0_distr_n[[index]]);abline(a= 0, b=1)
# use "xlim = c(0,40), ylim=c(0,40)" as arguents of "qqplot"
# to zoom into the tails of the qq plot
```

```
# Plot of the distribution of u0 approaching the
# chi-sq distribution as n increases
plot(x.grid,chisqcurve, type = 'l',
     main = "Distribution of u as n increases",
     xlab="u'", ylab="density",ylim = c(0,0.04), xlim = c(20,110))
for (i in 1:length(u0_distr_n)) {
  lines(density(u0_distr_n[[i]]),col = i+1)
  legend("topright", col=1:7, lty=1, cex=0.6
         legend = c("chi-sq","n=11","n=13","n=15","n=17","n=19","n=21"))
}

# Plots the distance from the chi-sq distribution
u0_distance <- list()
for (i in 1:length(u0_distr_n)){
  u0_distance[[i]] <- sum(abs(
            quantile(density(u0_distr_n[[i]])$x,seq(0,0.99,0.01))
            -qchisq(seq(0,0.99,0.01),degfreed)))/100
}
plot(num_of_obs,u0_distance,type = "o",ylim=c(0,15000))


# Plots the value alpha as n increases
plot(num_of_obs,u0_alpha_n, type = "o");abline(h = 0.05)


#### Plots comparing u0 and u1 ####
index = 1

plot(num_of_obs,u1_distance,type = "o", xlab="n", ylab="distance",
     ylim=c(0,140),
     main = "'Distance' of u and u' from chi-sq as n increases")
lines(num_of_obs,u0_distance,type = "o",col = 'red');abline(h=0.05,lty=2)
legend("topright", col = 1:2, legend = c("u'","u"),lty=1)

plot(num_of_obs,u1_alpha_n,type = "o",ylim=c(0,1), xlab="n", ylab="alpha",
     main = "Type I error of u and u' as n increases")
lines(num_of_obs,u0_alpha_n,type = "o",col = 'red');abline(h=0.05,lty=2)
legend("topright", col = 1:2, legend = c("u'","u"),lty=1)

plot(x.grid,chisqcurve, type = 'l')
lines(density(u1_distr_n[[index]]),col = 'red')
lines(density(u0_distr_n[[index]]),col = 'red')

par(mfrow=c(6,2))
par(oma=c(5,5,5,5))
par(mar=c(0.2,0.2,0.2,0.2))
par(mgp=c(2.5,1,0))

for (i in 1:6) {
  hist(u0_distr_n[[i]], prob = TRUE, col = "white",
     ylim = c(0, 0.05), xlim=c(10,110),
     breaks = 50, ylab = "", main = "", xlab = "",
     xaxt = "n", yaxt = "n")
```

```
  lines(x.grid, chisqcurve, col = 2, lwd = 2)
  hist(u1_distr_n[[i]], prob = TRUE, col = "white",
     ylim = c(0, 0.05), xlim=c(10,110),
     breaks = 50, ylab = "", main = "", xlab = "",
     xaxt = "n", yaxt = "n")
  lines(x.grid, chisqcurve, col = 2, lwd = 2)
}

mtext(paste(c('u', "u'")),at=c(0.3,0.8),side=3,outer=T,line = 1)
mtext(paste('n =',rev(num_of_obs)),at=seq(0.1,0.9,0.9/6),
            side=2,outer=T,line = 1)
```

A.2. R Code for Challenge 2

The code used to run the simulation is the following.

```
j=1
u0_power_rho_n <- list()
u1_power_rho_n <- list()

num_of_obs <- seq(from=11, to=100, by=10)
values_of_rho <- seq(from=-0.1, to=0.25, by=0.35/20)

for(N in num_of_obs) {
  nboot <- 10000 # Number of bootstraps
  alpha <- 0.05 # Significance level
  P <- 10 # Number of variables
  degfreed <- 1/2 * P * (P+1) -1

  u0_distr_rho <- list()
  u0_power_rho <- list()
  u1_distr_rho <- list()
  u1_power_rho <- list()

  k = 1

  quantile_u0 <- qchisq(1-alpha,degfreed)
  quantile_u1 <- qchisq(1-alpha,degfreed)


  for(rho in values_of_rho) {
    # Parameters for multivariate normal distribution,
    # under a Compound Symmetry covariance structure
    mu <- rep(0,P) # Mean
    sigma <- matrix(rho,P,P) - (rho - 1) * diag(P) # Covariance matrix

    if(is.positive.definite(sigma)==TRUE) {
      u0_distr <- rep(0,nboot)
      u1_distr <- rep(0,nboot)


      for (i in 1:nboot) {
        mvn <- mvrnorm(N, mu = mu, Sigma = sigma ) # from MASS package
```

```
        u0_distr[i] <- u0(mvn)
        u1_distr[i] <- u1(mvn)

      }

      u0_distr_rho[[k]] <-  u0_distr
      u0_power_rho[[k]] <- (length(which(u0_distr_rho[[k]] > quantile_u0)))
                            / length(u0_distr_rho[[k]])
      u1_distr_rho[[k]] <-  u1_distr
      u1_power_rho[[k]] <- (length(which(u1_distr_rho[[k]] > quantile_u1)))
                            / length(u1_distr_rho[[k]])
    } else {
      u0_power_rho[[k]] <- 1
      u1_power_rho[[k]] <- 1

    }

    inc(k)
  }
  u0_power_rho_n[[j]] <- u0_power_rho
  u1_power_rho_n[[j]] <- u1_power_rho

  inc(j)
}
```

Below is the code for the plots.

```
##### Plots for u0 (swap u0 with u1 to obtain the plots for u1)#####
plot(values_of_rho,u0_power_rho_n[[1]], type = "l",
    xlim = c(-0.1,0.1),ylim = c(0,1),
    main = "Power curve of u as n increases",
    xlab="rho", ylab="power")
for(i in 2:length(u0_power_rho_n)-1){
  lines(values_of_rho,u0_power_rho_n[[i]], type = "l",col = i)
  abline(h = 0.05,lty=2)
  legend("bottomleft", col = 1:(length(u0_power_rho_n)-1),
      lty=1,cex = 0.8,
      legend = c("n=11","n=21","n=31","n=41","n=51","n=61","n=71","n=81"))
}


#### Plots comparing u0 and u1 ####
par(mfrow=c(8,1))
par(oma=c(5,5,5,5))
par(mar=c(0.2,2,0.2,2))
par(mgp=c(2.5,0.1,0))

for (i in 1:8) {
  plot(values_of_rho,u1_power_rho_n[[i]], type = "l",
       xlim = c(-0.1,0.1),ylim = c(0,1),xaxt = "n")
  lines(values_of_rho,u0_power_rho_n[[i]], type = "l",col = 2)
  abline(h = 0.05, lty=2)
}
legend("bottomleft", col = 1:2, legend = c("u'","u"),lty=1,cex= 0.8)
```

```
#mtext(paste(c('u')),at=c(0.5),side=3,outer=T,line = 1)
mtext(paste('n =',rev(num_of_obs[1:8])),
      at=seq(0.05,0.95,1/8),side=2,outer=T,line = 1)
```

A.3. R Code for Challenge 3

The code used to run the simulation is the following.

```
u0_power_rho_n_pca <- list()
u1_power_rho_n_pca <- list()


for (l in 1:4) {
  j=1
  u0_power_rho_n <- list()
  u1_power_rho_n <- list()


  num_of_obs <- seq(from=11, to=100, by=20)
  values_of_rho <- seq(from=0, to=0.4, by=0.4/15)

  for(N in num_of_obs) {
    nboot <- 5000 # Number of bootstraps
    alpha <- 0.05 # Significance level
    P <- 10 # Number of variables
    degfreed <- 1/2 * P * (P+1) -1
    u0_distr_rho <- list()
    u0_power_rho <- list()
    u1_distr_rho <- list()
    u1_power_rho <- list()

    k = 1
    quantile_u0 <- qchisq(1-alpha,degfreed)
    quantile_u1 <- qchisq(1-alpha,degfreed)

    for(rho in values_of_rho) {
      # Parameters for multivariate normal distribution,
      # under a Compound Symmetry covariance structure
      mu <- rep(0,P) # Mean
      sigma <- matrix(rho,P,P) - (rho - 1) * diag(P) # Covariance matrix

      if(is.positive.definite(sigma)==TRUE) {
        u0_distr <- rep(0,nboot)
        u1_distr <- rep(0,nboot)


        for (i in 1:nboot) {
          mvn <- mvrnorm(N, mu = mu, Sigma = sigma ) # from MASS package
          mvn_pca <- prcomp(mvn)
          u0_distr[i] <- u0(mvn_pca$x[,1:(P+1-l)])
          u1_distr[i] <- u1(mvn_pca$x[,1:(P+1-l)])

        }
```

```
        u0_distr_rho[[k]] <-  u0_distr
        u0_power_rho[[k]] <- (length(which(u0_distr_rho[[k]] >
                                quantile_u0))) /
                             length(u0_distr_rho[[k]])
        u1_distr_rho[[k]] <-  u1_distr
        u1_power_rho[[k]] <- (length(which(u1_distr_rho[[k]] >
                                quantile_u1))) /
                             length(u1_distr_rho[[k]])


      } else {
        u0_power_rho[[k]] <- 1
        u1_power_rho[[k]] <- 1
      }

      inc(k)
    }
    u0_power_rho_n[[j]] <- u0_power_rho
    u1_power_rho_n[[j]] <- u1_power_rho

    inc(j)
  }

  u0_power_rho_n_pca[[l]] <- u0_power_rho_n
  u1_power_rho_n_pca[[l]] <- u1_power_rho_n
}
```

Below is the code for the plots.

```
##### Plots for u0 (swap u0 with u1 to obtain the plots for u1)#####
q = 5
plot(values_of_rho,u0_power_rho_n_pca[[1]][[q]], type = "l",
     xlim = c(-0.1,0.3), ylim = c(0,1))
for(i in 2:length(u0_power_rho_n_pca)){
  lines(values_of_rho,u0_power_rho_n_pca[[i]][[q]], type = "l",col = i)
}


#### Plots comparing u0 and u1 ####
par(mfrow=c(5,2))
par(oma=c(5,5,5,5))
par(mar=c(0.2,2,0.2,0.2))
par(mgp=c(2.5,1,0))

for (q in 1:5) {
    plot(values_of_rho,u0_power_rho_n_pca[[1]][[q]], type = "l",
         xlim = c(-0.1,0.3), ylim = c(0,1),
         ylab = "", xlab = "", xaxt = "n")
    for(i in 2:length(u0_power_rho_n_pca)){
      lines(values_of_rho,u0_power_rho_n_pca[[i]][[q]],
      type = "l",col = i)
    }
    plot(values_of_rho,u1_power_rho_n_pca[[1]][[q]], type = "l",
        yaxt = "n", xlim = c(-0.1,0.3), ylim = c(0,1),
```

```
        ylab = "", xlab = "", xaxt = "n")
    for(i in 2:length(u1_power_rho_n_pca)){
      lines(values_of_rho,u1_power_rho_n_pca[[i]][[q]],
      type = "l",col = i)
    }
}

mtext(paste(c('u', "u'")),at=c(0.25,0.75),
    side=3,outer=T,line = 1)
mtext(paste('n =',rev(num_of_obs)),
    at=seq(0.1,0.9,1/5),side=2,outer=T,line = 1)

plot(1,1)
legend("topright", col = 1:(length(u0_power_rho_n_pca)), lty=1
       legend = c("no variables removed",
                  "1 variable removed",
                  "2 variables removed",
                  "3 variables removed"))
```