Robust normality test and robust power transformation with application to state change detection in non normal processes

Dissertation

zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften
der Fakultät Statistik
der Technische Universität Dortmund

vorgelegt von

Arsene Ntiwa Foudjo

Gutachter
Prof. Dr. Roland Fried
Prof. Dr. Walter Krämer

Vorsitzenderin
Prof. Dr. Christine Müller

Tag der mündlichen Prüfung
04. Juni 2013

Contents

List of Figures

List of Tables

Motivation

Today more than ever before, we monitor time series and want to detect state changes in them. A state change can be defined as a change in the normal behaviour of the majority of the data. A common state change is for example the outlier. An outlier is an observation that is unusually far from the bulk of the data. Its effects on statistical procedures can be very severe if not appropriately dealt with. Correctly determining when the state changes occurs helps decision makers understand them and allows for a shorter reaction time and the definition of an appropriate course of action. Due to the amount of data available nowadays and the advances made in computer science (processor speed for instance), statistics (more complex models) and in other relevant fields of science, the task of building a good monitoring procedure is a much more challenging task today. Harrison & Stevens (1971) developed a state change detection procedure and a part of this thesis is based on their work. The main objective of this thesis is to improve this state change detection procedure by widening its spectrum of application and using up-to-date statistical techniques to enhance its results. One of the most frequent assumptions of statistical procedures is the one of normally distributed data. However, this requirement is usually not met in practice, for example manufacturing data are often positive and right skewed. Harrison & Stevens also make this assumption hereby narrowing the area of application of their procedure.

We opt for a transformation of the data to achieve approximate normality. From the wide range of transformation procedures, we chose to use the Box-Cox transformation. The parameter of the transformation must be estimated from a start sequence of the data if historical data is not available. A particular challenge are the outliers that can occur in the start sequence, so that we need a robust estimator of the transformation parameter. Several robust estimators are available, but the problem is that the transformation only achieves approximate normality. Therefore we define a new robust estimator which is based on the optimization of a new robust measure of normality. This idea is further supported by the fact that there is actually not only one transformation that can yield approximate normality, or there could be no transformation that achieves normality at all.

Speaking of measures of normality directs us towards tests for normality. Several robust test for normality have been developed but to the best of our knowledge, none of them is a robustification of the Shapiro-Wilk test which is known to be one of the most powerful tests for normality and has already been used to estimate the Box-Cox transformation parameter non robustly.

In Chapter 1, we derive a robust Shapiro-Wilk test for normality and determine its asymptotic null distribution. Simulations show that our new robust test outperforms its competitors in many respects, as expected. In Chapter 2, we use the robust Shapiro-Wilk test statistic to derive a robust estimator of the Box-Cox transformation parameter. The new robust estimator outperforms all the other robust estimators and the maximum-likelihood estimator in terms of better transformation to normality and bias in presence of outliers, but yields a slightly larger variance than its competitors. Finally we apply the robust transformation in Chapter 3 to transform the data before conducting the state change detection procedure. After some other improvements to the procedure, the obtained results are very appealing and most of our goals have been reached. We conclude this work with a summary and an outlook.

Chapter 1

Robust Shapiro-Wilk test for normality

1.1 Introduction

A large number of statistical methods relies on the assumption of normality, but in practice, this assumption is not always met. In this case, statistical procedures based on normality can suffer drastic consequences. A wide range of statistical procedures have been developed to test for normality. The most common tests for normality are the Shapiro-Wilk test and the Jarque-Bera test, also known as Browman-Shenton test. These tests are powerful but also very sensible to outliers. However, many statistical procedures based on robust estimators or outlier detection and elimination perform well if the data is normal with some outliers. Before applying such procedures, one wants to check whether the majority of the data comes from a normal distribution, or whether the data is not normal at all. In the presence of outliers, one thus needs robust tests. Robustifications of the Jarque-Bera test have been proposed by Brys et al. (2004b) and by Gel & Gastwirth (2008). Gel et al. (2007) introduced a new robust normality test against heavy-tailed alternatives. In this chapter, we propose a robust version of the Shapiro-Wilk test of normality. In Section 2, we present some classical tests of normality. In Section 3, we review some robust tests for normality found in the literature. We introduce our new robust test in Section 4 and derive some asymptotic properties in Section 5. We compare the already existing robust tests to our proposal in Section 6 via simulations.

1.2 Tests of normality

1.2.1 The Shapiro-Wilk test

Let $Y = (y_1, y_2, \ldots, y_n)^t$ be n ordered observations of iid data. The Shapiro-Wilk test is used to test the given data for normality. For this purpose, let $m = (m_1, m_2, \ldots, m_n)^t$ denote a vector of expected values of a standard normal order statistic and let $V_0 = (v_{ij})$ be the corresponding $n \times n$ covariance matrix. This means, if $X_{n1} \leq X_{n2} \leq \cdots \leq X_{nn}$ is an ordered statistic from n iid standard normal random variables, then the following holds:

$$\begin{aligned} E(X_{ni}) &= m_i, \quad i = 1, 2, \ldots, n\,, \\ Cov(X_{ni}, X_{nj}) &= v_{ij}, \quad i, j = 1, 2, \ldots, n\,. \end{aligned}$$

The W statistic of the Shapiro-Wilk test is given by

$$W = \frac{\left(\sum_{i=1}^{n} a_i y_i\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2}\,,$$

where

$$a^t = (a_1, a_2, \ldots, a_n) = \frac{m^t V_0^{-1}}{\left(m^t V_0^{-1} V_0^{-1} m\right)^{\frac{1}{2}}}\,. \tag{1.1}$$

The W statistic has the following analytical properties:

- W is location and scale invariant.
- The distribution of W depends only on the sample size n, for samples from a normal distribution.
- W is statistically independent from the sample variance and mean, for a sample from a normal distribution.
- The maximum value of W is 1 and the minimum is $\dfrac{na_1^2}{n-1}$.

- The half and first moments of W are given by:

$$EW^{1/2} = \frac{R^2\Gamma(\frac{n-1}{2})}{C\Gamma(\frac{n}{2})\sqrt{2}}$$
$$EW = \frac{R^2(R^2+1)}{C^2(n-1)},$$

  where $R^2 = m^t V_0^{-1} m$ and $C^2 = m^t V_0^{-1} V_0^{-1} m$.

- For $n = 3$, the density of W is

$$\frac{3}{\pi}(1-w)^{-1/2} w^{-1/2}, \quad 3/4 \leq w \leq 1.$$

The density of the W statistic is difficult to determine for sample sizes greater than 20. This difficulty arises from the determination of the values of the a_i given in (1.1) necessary for the computation of the W statistic, due to the fact that the covariance matrix $V_0 = (v_{ij})$ was only obtainable up to sample size 20. To solve the problem, Shapiro & Wilk (1965) proposed an approximation of the a_i and proved that when the sample size grows the difference between the approximated and the exact values tends to zero.

Royston (1995) gave an approximation to the weights a_i that can be used for any n in the range $3 \leq n \leq 5000$ and proposed an algorithm to perform the Shapiro-Wilk test.

Leslie et al. (1986) derived the asymptotic distribution of the Shapiro-Wilk test statistic and proved the consistency of the test using the work of Stephens (1975), who mainly shows that the vector m converges towards an eigenvector of V_0, i.e.

$$\|V_0 m - \frac{1}{2} m\| \xrightarrow{n\to\infty} 0,$$

where $\| \bullet \|$ is the Euclidean norm.

1.2.2 The Jarque-Bera test

The Jarque-Bera test of normality introduced in Jarque & Bera (1980) is based on the classical skewness and kurtosis coefficients denoted by γ_1 and γ_2, respectively, which are defined as follows:

$$\gamma_1(F) = \frac{\mu_3(F)}{\mu_2(F)^{3/2}} \tag{1.2}$$

$$\gamma_2(F) = \frac{\mu_4(F)}{\mu_2(F)^2}, \tag{1.3}$$

where F represents any distribution with finite k-th central moments $\mu_k \quad (k \leq 4)$. Under the normality assumption with $\gamma_1(F) = 0$ and $\gamma_2(F) = 3$ it is known that the sample skewness b_1 and the sample kurtosis b_2 are asymptotically independent and normally distributed (see Brys et al. (2004b)) with

$$\sqrt{n}\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \longrightarrow_{\mathcal{D}} \mathcal{N}\left(\begin{pmatrix} 0 \\ 3 \end{pmatrix}, \begin{pmatrix} 6 & 0 \\ 0 & 24 \end{pmatrix}\right). \tag{1.4}$$

This leads us to the Jarque-Bera test statistic

$$JB = n\left(\frac{b_1^2}{6} + \frac{(b_2-3)^2}{24}\right) \overset{H_0}{\longrightarrow}_{\mathcal{D}} \chi_2^2\,. \tag{1.5}$$

Because the sample skewness and kurtosis are sensible to outliers, this test is not robust to outliers.

1.3 Robust tests of normality in the literature

1.3.1 Robustification of the Jarque-Bera test

Brys et al. (2004b) proposed a robust version of the Jarque-Bera test, using the robust measure of skewness and tail weight of Brys et al. (2004a). Let $X^t = (x_1, x_2, \dots, x_n)$ be a sample of n independent observations from a continuous univariate distribution F. For simplicity, we assume $x_1 < x_2 < \cdots < x_n$. The medcouple (MC) is given by

$$MC = med_{x_i \leq m_n \leq x_j} h(x_i, x_j),$$

with $m_n = F_n^{-1}(0.5)$ the median of X_n and the kernel function h is defined for all $x_i \neq x_j$ as

$$h(x_i, x_j) = \frac{(x_j - m_n) - (m_n - x_i)}{x_j - x_i},$$

for some $i \neq j$.

If $x_i = x_j = m_n$ then there are observations tied to the median. Let the corresponding indices be $r_1 < r_2 < \cdots < r_k$ so that $x_{r_l} = m_n$ for all $l = 1, \ldots, k$. In this case, the kernel is given as follows:

$$h(x_{r_i}, x_{r_j}) = \begin{cases} -1, & \text{if } i+j-1 < k \\ 0, & \text{if } i+j-1 = k \\ +1, & \textit{if } i+j-1 > k \end{cases}$$

They also considered the left medcouple (LMC) and right medcouple (RMC), which are left and right tail weight measures, respectively. These robust measures are given by $LMC = -MC(x < m_n)$ and $RMC = MC(x > m_n)$. $MC(x > m_n)$ means that we compute the medcouple considering only observations in the sample that are larger than the median, and $MC(x < m_n)$ has an analogue meaning.

To define a general test statistic, let $\hat{\omega} = (\hat{\omega}_1, \hat{\omega}_2, \ldots, \hat{\omega}_k)$ be an estimator of an unknown parameter $\omega = (\omega_1, \omega_2, \ldots, \omega_k)$. If

$$\sqrt{n}(\hat{\omega}_1, \hat{\omega}_2, \ldots, \hat{\omega}_k) \longrightarrow_{\mathcal{D}} \mathcal{N}_k(\omega, \Sigma_k),$$

then the generalized test statistic $T = n(\hat{\omega} - \omega)^t \Sigma_k^{-1}(\hat{\omega} - \omega) \longrightarrow_{\mathcal{D}} \chi^2_k$. The Jarque-Bera test can be seen as a special case of this general test, if we take $k = 2$, $(\hat{\omega}_1, \hat{\omega}_2) = (b_1, b_2)$ and $(\omega_1, \omega_2) = (\gamma_1, \gamma_2) = (0, 3)$.

Brys et al. (2004b) proposed 3 alternative robust tests as special cases of the general test:

1. Test $MC1$ uses only the medcouple with $k = 1$, $\hat{\omega}_1 = MC$ and $\omega_1 = 0$.

2. Test $MC2$ uses $k = 2$, $(\hat{\omega}_1, \hat{\omega}_2) = (LMC, RMC)$ and $(\omega_1, \omega_2) = (0.199, 0.199)$.

3. Test $MC3$ uses $k = 3$, $(\hat{\omega}_1, \hat{\omega}_2, \hat{\omega}_3) = (MC, LMC, RMC)$ and $(\omega_1, \omega_2, \omega_3) = (0, 0.199, 0.199)$.

The asymptotic means ω_k mentioned above and the asymptotic covariance matrices Σ_k for each test are given in Brys et al. (2004b). They have been derived from the influence functions of the estimators.

The evaluation of the performances of the four tests of normality at Tukey's class of gh-distributions (see Hoaglin et al. (1985)) conducted in Brys et al. (2004b) leads them to the conclusion that the JB test outperforms the other tests in the absence of contamination, followed by $MC3$, which is much more conservative.

Additionally, contaminations were considered as mixtures of the form $(1-\delta)\mathcal{N}(0,1) + \delta\mathcal{N}(\mu,\sigma^2)$ where $\delta = 1\%$ or 5%. Samples of size 1000 and the following scenarios were considered:

- Scenario 1: $\mu = 0$ and $\sigma^2 = 0.05$
- Scenario 2: $\mu = 0$ and $\sigma^2 = 5$
- Scenario 3: $\mu = 7$ and $\sigma^2 = 1$
- Scenario 4: $\mu = -7$ and $\sigma^2 = 1$

The robust normality tests perform better than the JB test in all these contaminated cases. The performances of the robust tests are rather similar except in the last two scenarios, when $\delta = 5\%$. There the $MC3$ seems to be less robust than the other two robust tests.

1.3.2 Robust test of normality against heavy-tailed alternatives

Gel et al. (2007) introduced the SJ test as a new robust test of normality.

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with mean μ, median $\hat{\mu}_n$ and standard-deviation σ. The new robust SJ test is defined via the ratio

$$R = \frac{S_n}{J_n} \tag{1.6}$$

of the empirical standard deviation estimator S_n and a robust measure of spread J_n, the average absolute deviation from the sample median (MAAD) defined as

$$J_n = \frac{C}{n}\sum_{i=1}^{n} |x_i - \hat{\mu}_n|, \tag{1.7}$$

where $C = \sqrt{\pi/2}$.

The authors show that under the null hypothesis of normality, the following holds

$$\sqrt{n}(R-1) \longrightarrow_{\mathcal{D}} \mathcal{N}(0, \sigma_R^2), \tag{1.8}$$

with $\sigma_R^2 = \dfrac{\pi - 3}{2}$. This test is shown via simulations to outperform the Shapiro-Wilk test and the Jarque-Bera test with respect to power against heavy-tailed alternatives.

1.3.3 A robust modification of the Jarque-Bera test

Gel & Gastwirth (2008) have modified the Jarque-Bera test by replacing the sample skewness and kurtosis by more robust estimators. Let $X = (x_1, x_2, \ldots, x_n)^t$ be a sample of n independent observations with sample median $\hat{\mu}_n$. Let $\nu_k \quad (k \leq 4)$ be finite central moments and J_n the average absolute deviation from the sample median as defined in equation (1.7).

These authors define robust estimates of skewness and kurtosis as $\dfrac{\hat{\nu}_3}{J_n^3}$ and $\dfrac{\hat{\nu}_4}{J_n^4}$, respectively, where $\hat{\nu}_k$ denotes the sample estimate of the k-th central moment. Then the robust test statistic is defined as follows:

$$RJB = \frac{n}{C_1}\left(\frac{\hat{\nu}_3}{J_n^3}\right)^2 + \frac{n}{C_2}\left(\frac{\hat{\nu}_4}{J_n^4} - 3\right)^2, \tag{1.9}$$

where $C_1 = 6$ and $C_2 = 64$ are recommended by the authors to achieve the nominal significance level of $\alpha = 0.05$. Under the null hypothesis of normality the authors show that

$$\sqrt{n}\begin{pmatrix} \dfrac{\hat{\nu}_3}{J_n^3} \\ \dfrac{\hat{\nu}_4}{J_n^4} - 3 \end{pmatrix} \longrightarrow_{\mathcal{D}} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} C_1 & 0 \\ 0 & C_2 \end{pmatrix}\right), \tag{1.10}$$

so that RJB asymptotically follows a χ_2^2-distribution with 2 degrees of freedom. This test outperforms the Shapiro-Wilk, the Jarque-Bera and the SJ test (Gel et al., 2007) in terms of power against moderately heavy-tailed alternatives, especially in small and moderate sample sizes. The RJB is less powerful than the SJ test in case of heavy-tailed alternatives such as the double exponential distribution, because the SJ test is directed towards such alternatives.

1.3.4 Assessing when a sample is mostly normal

To robustify statistical procedures one often trims the data to exclude potential outliers. Alvarez-Esteban et al. (2010) present a method for trimming the data and afterwards comparing the distribution of the trimmed sample with a trimmed sample from a normal distribution to assess "mostly" normality.

They define the L_2-Wasserstein distance between two distributions P and Q on the real line as the L_2-distance between the quantile functions (respectively F^{-1} and G^{-1}) given by:

$$\mathcal{W}_2(P, Q) = \left[\int_0^1 \left(F^{-1}(t) - G^{-1}(t)\right)^2 dt\right]^{1/2} . \tag{1.11}$$

Let P be a probability measure on the real line and $0 \leq \alpha \leq 1$. P^* is called an α-trimming of P, if P^* is absolutely continuous with respect to P and $\frac{dP^*}{dP} \leq \frac{1}{1-\alpha}$. Define $\mathcal{T}_\alpha(P)$ as the set of α-trimmings of P. $\mathcal{T}_\alpha(P)$ can be parametrized in terms of α-trimmings of the uniform distribution on $(0, 1)$. Let $\mathcal{C}_\alpha$ be the class of absolutely continuous functions $h : [0, 1] \to [0, 1]$ such that $h(0) = 0$ and $h(1) = 1$, with derivative $0 \leq h' \leq \frac{1}{1-\alpha}$ and P_h denote the probability measure with distribution function $h(P(-\infty, t])$. Then we have:

$$\mathcal{T}_\alpha(P) = \{P_h : h \in \mathcal{C}_\alpha\} .$$

Assume that $X_1, \ldots, X_n$ are n i.i.d. observations with common distribution P and let $\mathcal{N}$ stand for the normal distribution family. Then the distance of the data sample from normality can be defined as follows:

$$\tau_\alpha(P, \mathcal{N}) = \inf_{h \in \mathcal{C}_\alpha, Q \in \mathcal{N}} \mathcal{W}_2^2 (P_h, Q_h) .$$

If $\tau_\alpha(P, \mathcal{N}) = 0$ then there is a normal distribution Q which is equal to P after removing a fraction of mass, of size at most α, from P and Q. A small value of $\tau_\alpha(P, \mathcal{N})$ indicates that most of the distribution underlying the data is not far from normality. Assessing "mostly" normality amounts to fixing a threshold Δ_0^2 and testing:

$$H_0 : \tau_\alpha(P, \mathcal{N}) \geq \Delta_0^2 \quad vs. \quad H_1 : \tau_\alpha(P, \mathcal{N}) < \Delta_0^2. \tag{1.12}$$

The authors show that $\tau_\alpha(P, \mathcal{N})$ is location invariant but not scale invariant. In order to solve this problem, they define

$$v(h) = \min_{Q \in \mathcal{N}} \mathcal{W}_2^2(P_h, Q_h)$$

so that $\tau_\alpha(P, \mathcal{N}) = \inf_{h \in C_\alpha} v(h)$. They assume that $v(h)$ admits a unique minimizer h_0 and define the standardized trimmed distance to normality as

$$\tilde{\tau}_\alpha(P, \mathcal{N}) = \frac{\tau_\alpha(P, \mathcal{N})}{Var(F^{-1} \circ h_0^{-1})},$$

which is scale invariant (Alvarez-Esteban et al., 2010), and they prove asymptotic normality of the new statistic. The new test problem is

$$H_0 : \tilde{\tau}_\alpha(P, \mathcal{N}) \geq \tilde{\Delta}_0^2 \quad vs. \quad H_1 : \tilde{\tau}_\alpha(P, \mathcal{N}) < \tilde{\Delta}_0^2, \tag{1.13}$$

where the thresholds Δ_0^2 and $\tilde{\Delta}_0^2$ from equations (1.12) and (1.13) are usually not the same.

In the next sections, when we run simulations and compare this test with the other robust tests, we shall refer to it with the name $TRIM_\alpha$.

1.4 New robust Shapiro-Wilk test

As we mentioned earlier, the Shapiro-Wilk test for normality can suffer severe consequences due to outliers. Since the Shapiro-Wilk test is one of the most powerful tests for normality, a robust version of this test can prove very useful. Our robustification method is based on detecting outliers, especially in skewed data. First we introduce some basic notions.

1.4.1 The adjusted boxplot

The boxplot is a well known and very useful tool in univariate data analysis. It gives us informations on the spread, the skewness and outliers, among other things. However, in the construction of the boxplot, we implicitly assume symmetry within the data since the same multiple of the interquartile range is used for the detection of upper and lower outliers. In addition, the position of the whiskers is determined by a rule assuming normality. The boxplot is therefore not well suited for the representation of skewed data. Hubert & Vandervieren (2008) developed an adjusted boxplot more suited for skewed datasets.

The usual boxplot consists of a box limited by the first and the third quartile. In addition, the whiskers are the lower and upper bounds determined so that under the normality assumption, data that falls outside the whiskers are considered to be outliers. The usual boundaries are given by

$$[Q_1 - 1.5IQR, Q_3 + 1.5IQR],$$

where Q_1 and Q_3 are respectively the first and the third quartile, and $IQR = Q_3 - Q_1$ denotes the interquartile range. For skewed data too many points fall outside the whiskers and are hence falsely classified as outliers.

The idea of the adjusted boxplot is to find a function $h(MC)$ of the medcouple (Brys et al., 2004b) to shift the whiskers and allow skewness in the data. After extensive simulations (Hubert & Vandervieren, 2008), the best boundaries were found to be

$$\begin{aligned}&[Q_1 - 1.5e^{-3.5MC}IQR, Q_3 + 1.5e^{4MC}IQR] \quad \text{when } MC \geq 0, \text{ and}\\ &[Q_1 - 1.5e^{-4MC}IQR, Q_3 + 1.5e^{3.5MC}IQR] \quad \text{otherwise.}\end{aligned}$$

The authors shows that the adjusted boxplot is much better suited for skewed data than the original one. Generally, it classifies less points falsely as outliers in case of skewed distributions. Note that the adjusted boxplot reduces to the original one in case of symmetrical data ($MC = 0$).

1.4.2 Outlier detection for skewed data

To detect outliers in skewed data, we can either use the adjusted boxplot introduced in the previous subsection or compute coefficients of outlyingness and derive a classification rule for outliers from them. Stahel (1981) and Donoho (1982) define the outlyingness of a univariate data point x_i within a sample $X = (x_1, x_2, \ldots, x_n)^t$ of length n as follows:

$$SDO_i = SDO(x_i, X) = \frac{|x_i - \hat{\mu}_n|}{\hat{\sigma}_n}, \tag{1.14}$$

where $\hat{\mu}_n$ is the sample median and $\hat{\sigma}_n$ is the median absolute deviation of the sample that we suppose to be corrected for consistency under normality.

This outlyingness coefficient measures the distance of each point to the center of the data standardized with a robust scale. It does not take skewness into account since it does not matter if the point is on the left or on the right hand side of the median. Hubert & Van der Veeken (2008) extended the outlyingness to an adjusted outlyingness defined by

$$AO_i = AO(x_i, X) = \begin{cases} \dfrac{x_i - \hat{\mu}_n}{w_1 - \hat{\mu}_n}, & \text{if } x_i \geq \hat{\mu}_n, \\ \dfrac{\hat{\mu}_n - x_i}{\hat{\mu}_n - w_2} & \text{otherwise,} \end{cases} \tag{1.15}$$

where w_1 and w_2 are the upper and lower whiskers of the adjusted boxplot applied to the dataset X introduced in the previous section. Note again that the adjusted outlyingness reduces to the outlyingness for symmetrical data. In theory, the adjusted outlyingness (AO) can resist 25% outliers, although one notices a substantial bias of the medcouple when the contamination exceeds 10%. Furthermore, the authors prove that the AO has a bounded influence function.

Under normality, the AO is asymptotically χ_1^2 distributed (for univariate data), but for skewed data the distribution is unknown. Hence to detect outliers after computing the AO, the authors propose to use the adjusted boxplot (AB) for right skewed data applied to the AO to classify points that are larger than the upper whiskers as outliers. We will use the adjusted boxplot and the adjusted outlyingness to detect outliers and then derive robust versions of the Shapiro-Wilk test in the same manner as our proposed robust normality test that we introduce in the next subsection. In the next sections, we shall refer to these tests respectively with the names RSW_{AB} and RSW_{AO}.

1.4.3 New robust tests of normality

Our idea is to apply an outlier detection procedure to derive a robust test for normality in the presence of outliers based on the Shapiro-Wilk test.

Let $X = (x_1, x_2, \ldots, x_n)^t$ denote a sample of ordered iid data of size n. Let $\hat{\mu}_n$ represent the sample median of X and $\hat{\sigma}_n$ the sample median absolute deviation (MAD), that are respectively robust estimators of the mean μ and the standard-deviation σ. We assume the MAD to be already corrected for consistency under the hypothesis of a normal distribution. Let $\tilde{X}$ be an additional ordered random sample of length n, generated artificially from the normal distribution with mean $\hat{\mu}_n$ and variance $\hat{\sigma}_n^2$.
We now investigate two new robustifications of the Shapiro-Wilk test.

1.4.3.1 Symmetrical trimming

Under the null hypothesis of normality, a large percentage of the data should be equally spread around the mean within a radius of about 3 standard deviations.
Let $\mathcal{O} = \{t | x_t \notin [\hat{\mu}_n - 3\hat{\sigma}_n, \hat{\mu}_n + 3\hat{\sigma}_n]\}$ denote the set of indices of assumed very unlikely (outlying) observations under the normality assumption. We define:

$$y_t = \begin{cases} x_t, & \text{if } t \notin \mathcal{O} \\ \tilde{x}_t, & \text{if } t \in \mathcal{O}. \end{cases}$$

The observations of the sample X are replaced by those of the sample $\tilde{X}$, so that the L_n smallest outlying observations in X are replaced by the L_n smallest observations of $\tilde{X}$, where L_n denotes the number of observations of X that are less than $\hat{\mu}_n - 3\hat{\sigma}_n$. The similar replacement procedure holds for the U_n largest observations of X, where by analogy U_n denotes the number of observations of X that are greater than $\hat{\mu}_n + 3\hat{\sigma}_n$. To test X for normality, we apply the Shapiro-Wilk test of normality to the modified data Y and denote the new test by RSW.

1.4.3.2 Asymmetrical trimming

Considering the fact that in the presence of skewed data it might make more sense to trim the data asymmetrically, we consider the left and right MAD that we define as:

$$\hat{\sigma}_{n,l} = MAD_{left} = c_0 \quad \underset{x_i < \hat{\mu}_n}{med}(\hat{\mu}_n - x_i)$$
$$\hat{\sigma}_{n,r} = MAD_{right} = c_0 \quad \underset{x_i > \hat{\mu}_n}{med}(x_i - \hat{\mu}_n),$$

where c_0 is a correction constant to achieve consistency under normality.
Analogously to the previous subsection, we consider $\mathcal{O} = \{t | x_t \notin [\hat{\mu}_n - 3\hat{\sigma}_{n,l}, \hat{\mu}_n + 3\hat{\sigma}_{n,r}]\}$ the set of indices of outlying observations under the normality assumption and define

$$y_t = \begin{cases} x_t, & \text{if } t \notin \mathcal{O} \\ \tilde{x}_t, & \text{if } t \in \mathcal{O}, \end{cases}$$

i.e. we replace the observations in the sample X in the same manner as we did in the symmetric case.
To test X for normality, we apply the Shapiro-Wilk test of normality to the modified data set Y. We call this test asymmetrical robustified Shapiro-Wilk test and denote it by RSW_{AS}.

1.5 Asymptotic theory for the new test

In this section, we investigate the asymptotic limit of the empirical distribution function of the symmetrically modified sequence and determine the asymptotic distribution of the new robust Shapiro-Wilk test statistic.

Let $X = (X_1, X_2, \ldots, X_n)^t$ be a sample of iid data from a normal distribution with mean μ and variance σ^2. Let $X_{(n)} = (X_{n1}, X_{n2}, \ldots, X_{nn})$ be the order statistic of the sample. We assume that each random variable X_i is defined on the probability space $(\Omega_1, \mathbb{A}_1, P)$, so that $X_i : \Omega_1 \longrightarrow \mathbb{R}$.

Further we define $\hat{\mu}_n$ and $\hat{\sigma}_n$ the sample median and the sample MAD as robust estimators for the mean and standard-deviation of X, i.e. these estimators are random variables defined on Ω_1.

Conditional on $\hat{\mu}_n$ and $\hat{\sigma}_n$, let $\tilde{X} = (\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n)^t$ be an artificially generated iid sample of size n from a normal distribution with mean $\hat{\mu}_n$ and variance $\hat{\sigma}_n^2$, so that $\tilde{X}_{(n)} = (\tilde{X}_{n1}, \tilde{X}_{n2}, \ldots, \tilde{X}_{nn})^t$ is the corresponding order statistic. Each random variable $\tilde{X}_i$ is defined on the probability space $(\Omega_1 \times \Omega_2, \mathbb{A}_1 \times \mathbb{A}_2, P \otimes P)$, so that $\tilde{X}_i : \Omega_1 \times \Omega_2 \longrightarrow \mathbb{R}$. For convenience, we will use P as probability measure for events in Ω_1 and $\Omega_1 \times \Omega_2$, when no confusion is possible.

We symmetrically trim X and use the values in $\tilde{X}$ to replace the outlying values of X, thus obtaining the sample $Y = (Y_1, Y_2, \ldots, Y_n)^t$ as explained above.

1.5.1 Notations

Let

- L_n denote the number of observations of X smaller than $\hat{\mu}_n - 3\hat{\sigma}_n$.
- U_n denote the number of observations of X larger than $\hat{\mu}_n + 3\hat{\sigma}_n$.
- Φ be the distribution function of the standard normal distribution.
- F be the distribution function of the normal distribution with mean μ and variance σ^2.

- $\tilde{F}_n$ be the empirical distribution function of the modified sequence Y.
- F_n denote the empirical distribution function of the original sequence X.
- G_n be the distribution function of the normal distribution with mean $\hat{\mu}_n$ and variance $\hat{\sigma}^2_n$.
- $\hat{G}_n$ be the empirical distribution function of the sequence $\tilde{X}$.
- $H_n(x)$ denote the number of observations among the L_n smallest and the U_n largest observations of $\tilde{X}$ that are less than x.
- $\mathbb{N}_n = \{i \in \mathbb{N}, \quad \text{so that} \quad i \leq n\}$ for $n \in \mathbb{N}$.

1.5.2 Asymptotic distribution of the modified sequence

The first result we want to establish is the asymptotic distribution of the modified sequence under the null hypothesis of normal distributed data.

Theorem 1

Assuming the original data X follows a normal distribution, we have

$$\sup_{x\in\mathbb{R}} \left|\tilde{F}_n(x) - F(x)\right| \xrightarrow{wp1} 0,$$

i.e. the empirical distribution of the modified sequence converges uniformly to the distribution function underlying the original data.

For the proof of Theorem 1 some auxiliary results are needed.

1.5.2.1 Properties

Lemma 1

The distribution function G_n of the normal distribution with mean $\hat{\mu}_n$ and variance $\hat{\sigma}_n^2$ fulfils

$$\sup_{x\in\mathbb{R}} |G_n(x) - F(x)| \xrightarrow{wp1} 0,$$

i.e. the distribution function from which the artificial data are generated converges uniformly to the distribution function underlying the original data.

Proof

We know that $\hat{\mu}_n \xrightarrow{wp1} \mu$ and $\hat{\sigma}_n \xrightarrow{wp1} \sigma$, see Serfling & Mazumder (2009), so that we have

$$\hat{\mu}_n \xrightarrow{wp1} \mu \quad \Rightarrow x - \hat{\mu}_n \xrightarrow{wp1} x - \mu \quad \forall x \in \mathbb{R} \quad \Rightarrow \frac{x - \hat{\mu}_n}{\hat{\sigma}_n} \xrightarrow{wp1} \frac{x-\mu}{\sigma} \quad \forall x \in \mathbb{R}.$$

Since the normal distribution function Φ is continuous, we can apply the continuous mapping theorem to obtain the following relation:

$$\begin{aligned} \Phi\left(\frac{x - \hat{\mu}_n}{\hat{\sigma}_n}\right) &\xrightarrow{wp1} \Phi\left(\frac{x-\mu}{\sigma}\right) \quad \forall x \in \mathbb{R} \\ \Rightarrow \left|\Phi\left(\frac{x - \hat{\mu}_n}{\hat{\sigma}_n}\right) - \Phi\left(\frac{x-\mu}{\sigma}\right)\right| &\xrightarrow{wp1} 0 \quad \forall x \in \mathbb{R} \end{aligned}$$

Because F is continuous, we can apply the theorem of Polya (Serfling, 1980, page. 18). Hence from the pointwise convergence follows the uniform convergence. $\blacksquare$

Lemma 2

Let $(\epsilon, x) \in \mathbb{R}^2$ be a vector with positive coordinates and $x + \epsilon < 1$. For the function $g_1(\epsilon, x)$ defined by

$$g_1(\epsilon, x) = \left(\frac{x}{x+\epsilon}\right)^{x+\epsilon} \left(\frac{1-x}{1-x-\epsilon}\right)^{1-x-\epsilon},$$

it follows that $0 < g_1(\epsilon, x) < 1$.

Proof

Define

$$\begin{aligned} g_2(\epsilon, x) &= \log\left(g_1(\epsilon, x)\right) \\ &= (x+\epsilon)\left[\log(x) - \log(x+\epsilon)\right] + (1-x-\epsilon)\left[\log(1-x) - \log(1-x-\epsilon)\right] . \end{aligned}$$

We have

$$\begin{aligned} \frac{\partial g_2}{\partial \epsilon}(\epsilon, x) &= \log\left(\frac{x}{x+\epsilon}\right) + (x+\epsilon)\left(\frac{-1}{x+\epsilon}\right) - \log\left(\frac{1-x}{1-x-\epsilon}\right) + (1-x-\epsilon)\left(\frac{1}{1-x-\epsilon}\right) \\ &= \log\left(\frac{x}{x+\epsilon}\right) + \log\left(\frac{1-x-\epsilon}{1-x}\right) < 0 \quad \forall \epsilon > 0 . \end{aligned}$$

This means that $g_2(\epsilon, x)$ is a monotone decreasing function of ϵ for every fixed value of x. It follows for all (ϵ, x) as defined above that

$$\begin{aligned} &\log\left(g_1(\epsilon, x)\right) = g_2(\epsilon, x) < g_2(0, x) = 0 \\ &\Leftrightarrow 0 < g_1(\epsilon, x) < 1 \quad \blacksquare \end{aligned}$$

Lemma 3

$$\sup_{x \in \mathbb{R}} \left| \hat{G}_n(x) - F(x) \right| \xrightarrow{wp1} 0,$$

i.e. the empirical distribution of the simulated data converges uniformly to the distribution function underlying the original data.

Proof

Let the sample $X_1, X_2, \ldots, X_n$ denote n independent realisations of the random variable $X : \Omega_1 \longrightarrow \mathbb{R}$ and for each $\omega_1 \in \Omega_1$, we generate artificially the sequence $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$ from a $\mathcal{N}(\hat{\mu}_n(\omega_1), \hat{\sigma}_n^2(\omega_1))$-distribution, representing n realisations of the random variable $\tilde{X} : \Omega_1 \times \Omega_2 \longrightarrow \mathbb{R}$. Furthermore let $x \in \mathbb{R}$.

Define for each $n \in \mathbb{N}$ and $\omega_1 \in \Omega_1$, $Z_{ni}(\omega_1, \bullet) = I\left(\tilde{X}_{ni}(\omega_1, \bullet) \leq x\right)$ for $\forall i \in \mathbb{N}_n$. For each fixed value of $n \in \mathbb{N}$, and fixed value of $\omega_1 \in \Omega_1$, the sequence $Z_{n1}, Z_{n2}, \ldots$ is an iid sequence of Bernoulli distributed data with parameter $\alpha_n(\omega_1) = G_n(x)(\omega_1)$.

According to Lemma 1, we have $\alpha_n \xrightarrow{wp1} F(x)$. Let in the following

$$A = \left\{\omega_1 \in \Omega_1 : \alpha_n(\omega_1) \xrightarrow{n\to\infty} F(x)\right\}.$$

We have $P(A) = 1$.

Fix $\omega_1 \in A$ and $\epsilon > 0$ at an arbitrary value. Define

$$\begin{aligned} P_n(\omega_1) &= P\left(\left|\frac{1}{n}\sum_{j=1}^{n} Z_{nj}(\omega_1, \bullet) - \alpha_n\right| \geq \epsilon\right) = P_{n1}(\omega_1) + P_{n2}(\omega_1) \quad \text{where} \\ P_{n1}(\omega_1) &= P\left(\frac{1}{n}\sum_{j=1}^{n} Z_{nj}(\omega_1, \bullet) \geq \alpha_n + \epsilon\right) \quad \text{and} \\ P_{n2}(\omega_1) &= P\left(\frac{1}{n}\sum_{j=1}^{n} Z_{nj}(\omega_1, \bullet) \leq \alpha_n - \epsilon\right). \end{aligned}$$

We now show that

$$\sum_{n=1}^{\infty} P_n(\omega_1) < \infty.$$

It is clear that if $\epsilon \geq 1$, then we have $P_n(\omega_1) = P_{n1}(\omega_1) = P_{n2}(\omega_1) = 0$. Hence we will consider $\epsilon < 1$ in the remaining of the proof. Let us study the two probabilities $P_{n1}(\omega_1)$ and $P_{n2}(\omega_1)$ for a fixed value of $n \in \mathbb{N}$, starting with the case of $P_{n1}(\omega_1)$.

If $\alpha_n(\omega_1) + \epsilon > 1$ holds, we have $P_{n1}(\omega_1) = 0$.

If $\alpha_n(\omega_1) + \epsilon = 1$, we get $P_{n1}(\omega_1) = P\left(\frac{1}{n}\sum_{j=1}^{n} Z_{nj}(\omega_1, \bullet) \geq 1\right) = (\alpha_n(\omega_1))^n$.

If $\alpha_n(\omega_1) + \epsilon < 1$ holds, Chernoff (1952) and Hoeffding (1963) show that we can bound

$P_{n1}(\omega_1)$ as follows

$$P_{n1}(\omega_1) \leq \left[\left(\frac{\alpha_n(\omega_1)}{\alpha_n(\omega_1) + \epsilon} \right)^{\alpha_n(\omega_1)+\epsilon} \left(\frac{1 - \alpha_n(\omega_1)}{1 - \alpha_n(\omega_1) - \epsilon} \right)^{1-\alpha_n(\omega_1)-\epsilon} \right]^n .$$

Let

$$g_1(\epsilon, \alpha_n(\omega_1)) = \left(\frac{\alpha_n(\omega_1)}{\alpha_n(\omega_1) + \epsilon} \right)^{\alpha_n(\omega_1)+\epsilon} \left(\frac{1 - \alpha_n(\omega_1)}{1 - \alpha_n(\omega_1) - \epsilon} \right)^{1-\alpha_n(\omega_1)-\epsilon} .$$

Since $\alpha_n(\omega_1) + \epsilon < 1$ holds, we have $0 < g_1(\epsilon, \alpha_n(\omega_1)) < 1$ (see Lemma 2) . Figure 1.1 shows a surface plot of the function $g_1(\epsilon, \alpha_n)$.

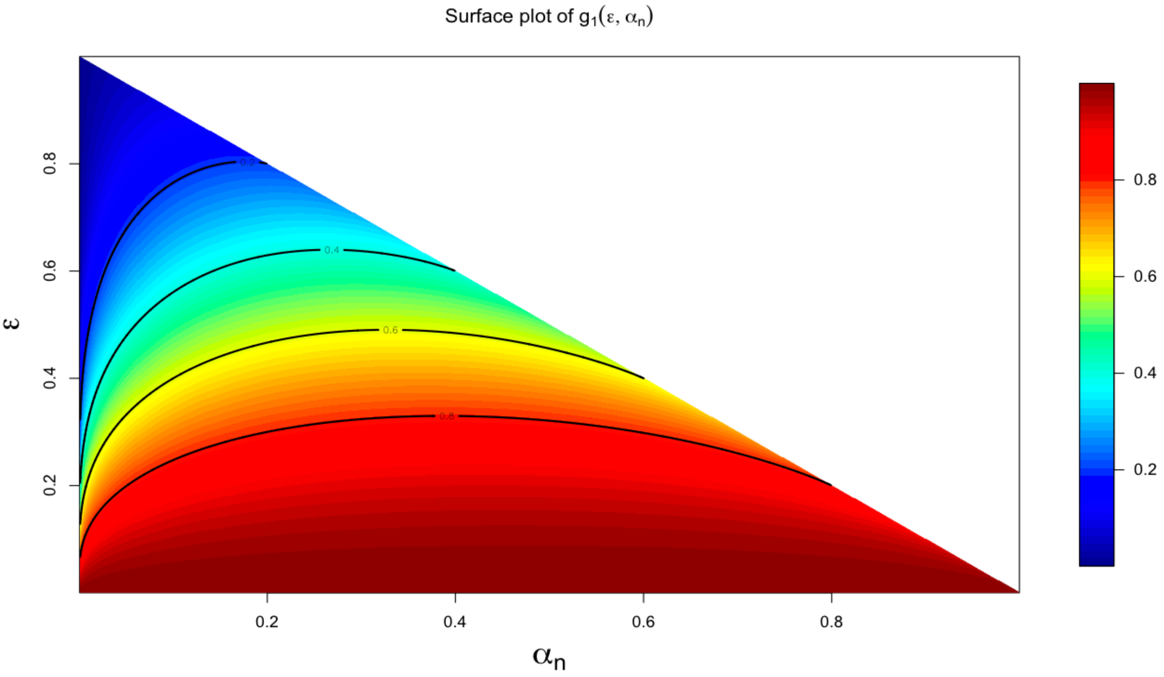


Figure 1.1: Surface plot of $g_1(\epsilon,\alpha_n)$.

Also note that for every fixed value of ϵ, $g_1(\epsilon, \alpha_n(\omega_1))$ is a continuous function of $\alpha_n(\omega_1)$. From the convergence of $\alpha_n(\omega_1)$ in Lemma 1, it follows that $\alpha_n(\omega_1)$ is bounded in a closed set $\Theta(\omega_1) \subset (0, 1)$ for n sufficiently large. Since $g_1(\epsilon, \alpha_n(\omega_1)) \in (0, 1)$ for any finite $n \in \mathbb{N}$, it follows that $\exists \theta_\epsilon(\omega_1) \in (0, 1)$ such that

$$\theta_\epsilon(\omega_1) = \sup_{n \in \mathbb{N}} g_1(\epsilon, \alpha_n(\omega_1)) .$$

This yields

$$P_{n1}(\omega_1) \leq \theta_\epsilon^n(\omega_1)\,.$$

Hence we have shown that

$$\begin{cases} P_{n1}(\omega_1) = 0, & \text{if } \alpha_n(\omega_1) + \epsilon > 1 \\ P_{n1}(\omega_1) = \alpha_n^n(\omega_1), & \text{if } \alpha_n(\omega_1) + \epsilon = 1 \\ P_{n1}(\omega_1) \leq \theta_\epsilon^n(\omega_1), & \text{if } \alpha_n(\omega_1) + \epsilon < 1 \end{cases}$$

In conclusion, if we define $b_\epsilon(\omega_1) = \sup_{n\in\mathbb{N}}(\alpha_n(\omega_1), \theta_\epsilon(\omega_1))$, we get that $b_\epsilon(\omega_1) < 1$ and $P_{n1}(\omega_1) \leq b_\epsilon^n(\omega_1)$.

By analogy to the case of $P_{n1}(\omega_1)$, we can show also by using the Chernoff inequality for $P_{n2}(\omega_1)$ that $P_{n2}(\omega_1) \leq a_\epsilon^n(\omega_1)$, for some $a_\epsilon(\omega_1) \in (0,1)$. This implies that

$$\begin{aligned} &\forall\epsilon > 0: \quad P_n(\omega_1) \leq b_\epsilon^n(\omega_1) + a_\epsilon^n(\omega_1) \\ \Rightarrow &\forall\epsilon > 0: \quad \sum_{n=1}^{\infty} P_n(\omega_1) \leq \sum_{n=1}^{\infty} b_\epsilon^n(\omega_1) + \sum_{n=1}^{\infty} a_\epsilon^n(\omega_1) \\ \Leftrightarrow &\forall\epsilon > 0: \quad \sum_{n=1}^{\infty} P\left(\left|\frac{1}{n}\sum_{j=1}^{n} Z_{nj}(\omega_1, \bullet) - \alpha_n(\omega_1)\right| \geq \epsilon\right) \leq \frac{1}{1-b_\epsilon(\omega_1)} + \frac{1}{1-a_\epsilon(\omega_1)} < \infty\,. \end{aligned}$$

We have shown that

$$\forall\epsilon > 0: \quad \sum_{n=1}^{\infty} P\left(\left|\frac{1}{n}\sum_{j=1}^{n} Z_{nj}(\omega_1, \bullet) - \alpha_n(\omega_1)\right| \geq \epsilon\right) < \infty \quad \text{with probability one}\,.$$

The theorem of Borel-Cantelli thus yields

$$\begin{aligned} &\forall\omega_1 \in A: \quad \left|\frac{1}{n}\sum_{j=1}^{n} Z_{nj}(\omega_1, \bullet) - \alpha_n(\omega_1)\right| \xrightarrow{wp1} 0 \\ \Leftrightarrow &\forall\omega_1 \in A: \quad \left|\frac{1}{n}\sum_{j=1}^{n} Z_{nj}(\omega_1, \bullet) - G_n(x)(\omega_1)\right| \xrightarrow{wp1} 0\,. \end{aligned}$$

Define the set

$$\begin{aligned} N &= \{\omega = (\omega_1, \omega_2) : \omega_1 \notin A \quad \text{or} \quad \omega_2 \in C(\omega_1)\}, \quad \text{where} \\ C(\omega_1) &= \{\omega_2 \in \Omega_2 : \left|\frac{1}{n}\sum_{j=1}^{n} Z_{nj}(\omega_1, \omega_2) - G_n(x)(\omega_1)\right| \nrightarrow 0\}\,. \end{aligned}$$

Fubini's theorem implies $P(N) = 0$. Since

$$\forall \omega = (\omega_1, \omega_2) \in N^c : \quad \left| \frac{1}{n} \sum_{j=1}^{n} Z_{nj}(\omega_1, \omega_2) - G_n(x)(\omega_1) \right| \to 0 \quad \forall x \in \mathbb{R}$$

$$\Leftrightarrow \left| \hat{G}_n(x) - G_n(x) \right| = \left| \frac{1}{n} \sum_{j=1}^{n} Z_{nj} - G_n(x) \right| \xrightarrow{wp1} 0 \quad \forall x \in \mathbb{R}$$

We get with the use of Lemma 1

$$\left| \hat{G}_n(x) - F(x) \right| \xrightarrow{wp1} 0, \quad \forall x \in \mathbb{R}\,.$$

Again with the theorem of Polya (Serfling, 1980, page. 18) follows the result

$$\sup_{x \in \mathbb{R}} \left| \hat{G}_n(x) - F(x) \right| \xrightarrow{wp1} 0\,. \quad \blacksquare$$

Lemma 4

The fraction $\dfrac{L_n}{n}$ *of replaced lower outliers fulfils*

$$\left| \frac{L_n}{n} - \Phi(-3) \right| \xrightarrow{wp1} 0\,.$$

Proof

From the theorem of Glivenko-Cantelli follows that

$$D_{1n} = \left| F_n\left(\hat{\mu}_n - 3\hat{\sigma}_n\right) - F\left(\hat{\mu}_n - 3\hat{\sigma}_n\right) \right| \leq \sup_{x \in \mathbb{R}} \left| F_n(x) - F(x) \right| \xrightarrow{wp1} 0\,.$$

Using the continuous mapping theorem, we obtain

$$\begin{aligned} \hat{\mu}_n - 3\hat{\sigma}_n &\xrightarrow{wp1} \mu - 3\sigma \\ \Rightarrow D_{2n} &= \left| F\left(\hat{\mu}_n - 3\hat{\sigma}_n\right) - F\left(\mu - 3\sigma\right) \right| \xrightarrow{wp1} 0\,. \end{aligned}$$

We deduce

$$\begin{aligned} \left| \frac{L_n}{n} - \Phi(-3) \right| &= \left| F_n\left(\hat{\mu}_n - 3\hat{\sigma}_n\right) - F\left(\mu - 3\sigma\right) \right| \\ &\leq D_{1n} + D_{2n} \xrightarrow{wp1} 0\,. \end{aligned}$$

Hence

$$\left| \frac{L_n}{n} - \Phi(-3) \right| \xrightarrow{wp1} 0\,. \quad \blacksquare$$

Lemma 5

The fraction $\frac{U_n}{n}$ *of replaced upper outliers fulfils*

$$\left|\frac{U_n}{n} - \Phi(-3)\right| \xrightarrow{wp1} 0\,.$$

Proof

The proof is analogue to that of Lemma 4. ▮

1.5.2.2 Proof of Theorem 1

Proof

Given $\hat{\mu}_n$, $\hat{\sigma}_n$, L_n and U_n, we can write:

$$\tilde{F}_n(x) = \begin{cases} \frac{H_n(x)}{n}, & \text{if } x < \hat{\mu}_n - 3\hat{\sigma}_n \\ F_n(x) - \frac{L_n}{n} + \frac{H_n(x)}{n}, & \text{if } \hat{\mu}_n - 3\hat{\sigma}_n \leq x \leq \hat{\mu}_n + 3\hat{\sigma}_n \\ \frac{n-U_n-L_n}{n} + \frac{H_n(x)}{n}, & \text{if } x > \hat{\mu}_n + 3\hat{\sigma}_n \end{cases} \tag{1.16}$$

This is equivalent to the following equation:

$$\tilde{F}_n(x) = \frac{H_n(x)}{n} + \min\left\{\max\left\{F_n(x) - \frac{L_n}{n}, 0\right\}, 1 - \frac{U_n + L_n}{n}\right\}, \tag{1.17}$$

where

$$\frac{H_n(x)}{n} = \min\left\{\hat{G}_n(x), \frac{L_n}{n}\right\} + \max\left\{\hat{G}_n(x) - \frac{n - U_n}{n}, 0\right\}\,. \tag{1.18}$$

We know that given two real numbers u and v, we have

$$\min\{u,v\} = \frac{1}{2}\left(u + v - |u - v|\right) = \min\{u - v, 0\} + v$$
$$\max\{u,v\} = \frac{1}{2}\left(u + v + |u - v|\right) = \max\{u - v, 0\} + v\,.$$

Hence $\min\{u, v\}$ and $\max\{u, v\}$ are both continuous functions of their arguments, so that we can apply the continuous mapping theorem to both functions. It follows with Lemmas 3, 4, and 5 and with equations (1.17) and (1.18) that we can write:

$$\forall x \in \mathbb{R}, \quad \tilde{F}_n(x) \xrightarrow{wp1} \tilde{F}(x) = \min\{F(x), \Phi(-3)\} + \max\{F(x) - (1 - \Phi(-3)), 0\} +$$
$$+ \min\{\max\{F(x) - \Phi(-3), 0\}, 1 - 2\Phi(-3)\}\,.$$

We consider three cases:

<u>Case 1</u>: $x < \mu - 3\sigma$

In this case, we have $F(x) < F(\mu - 3\sigma) = \Phi(-3)$, so that

$$\begin{aligned}\tilde{F}(x) =& \min\{F(x), \Phi(-3)\} + \max\{F(x) - (1 - \Phi(-3))\,, 0\} + \\ & + \min\{\max\{F(x) - \Phi(-3), 0\}, 1 - 2\Phi(-3)\} \\ =& F(x) + 0 + \min\{0, 1 - 2\Phi(-3)\} = F(x) + 0, \\ \text{i.e.} \quad \tilde{F}(x) =& F(x)\,.\end{aligned}$$

<u>Case 2</u>: $\mu - 3\sigma \leq x \leq \mu + 3\sigma$

Here, we have $\Phi(-3) = F(\mu - 3\sigma) \leq F(x) \leq F(\mu + 3\sigma) = \Phi(3)$. It follows

$$\begin{aligned}\tilde{F}(x) =& \min\{F(x), \Phi(-3)\} + \max\{F(x) - (1 - \Phi(-3))\,, 0\} + \\ & + \min\{\max\{F(x) - \Phi(-3), 0\}, 1 - 2\Phi(-3)\} \\ =& \Phi(-3) + 0 + \min\{F(x) - \Phi(-3), 1 - 2\Phi(-3)\} \\ =& \min\{F(x), 1 - \Phi(-3)\} = \min\{F(x), \Phi(3)\}, \\ \text{i.e.} \quad \tilde{F}(x) =& F(x)\,.\end{aligned}$$

<u>Case 3</u>: $x > \mu + 3\sigma$

Here, it holds $F(x) > F(\mu + 3\sigma) = \Phi(3)$ and it follows

$$\begin{aligned}\tilde{F}(x) =& \min\{F(x), \Phi(-3)\} + \max\{F(x) - (1 - \Phi(-3))\,, 0\} + \\ & + \min\{\max\{F(x) - \Phi(-3), 0\}, 1 - 2\Phi(-3)\} \\ =& \Phi(-3) + \max\{F(x) - \Phi(3), 0\} + \min\{F(x) - \Phi(-3), 1 - 2\Phi(-3)\} \\ =& \Phi(-3) + F(x) - \Phi(3) + \min\{F(x), 1 - \Phi(-3)\} - \Phi(-3) \\ =& F(x) - \Phi(3) + \min\{F(x), \Phi(3)\} = F(x) - \Phi(3) + \Phi(3), \\ \text{i.e.} \quad \tilde{F}(x) =& F(x)\,.\end{aligned}$$

In conclusion, we obtain $\forall x \in \mathbb{R}, \quad \tilde{F}(x) = F(x)$ so that with probability one

$$\forall x \in \mathbb{R}, \quad \lim_{n\to\infty} \left|\tilde{F}_n(x) - F(x)\right| = 0\,.$$

Thus, $\tilde{F}_n(x)$ converges pointwise to $F(x)$ for all $x \in \mathbb{R}$. Because F is continuous, we can apply the theorem of Polya (Serfling, 1980, page. 18). This yields the uniform convergence with probability one:

$$\lim_{n\to\infty} \left|\tilde{F}_n(x) - F(x)\right| = 0, \quad \forall x \in \mathbb{R}\,. \qquad \blacksquare$$

1.5.3 Asymptotic distribution of the new robust test statistic

We use the notations introduced in Section 1.5.1. Let us define the Shapiro-Wilk test statistic (Shapiro & Wilk, 1965) and the robust Shapiro-Wilk test statistic respectively as follows

$$W_n = \frac{\left(\sum_{k=1}^{n} a_{nk} X_{nk}\right)^2}{\sum_{k=1}^{n} \left(X_{nk} - \bar{X}\right)^2}$$

and

$$\tilde{W}_n = \frac{\left(\sum_{k=1}^{n} a_{nk} Y_{nk}\right)^2}{\sum_{k=1}^{n} \left(Y_{nk} - \bar{Y}\right)^2},$$

where $\bar{X}$ and $\bar{Y}$ are the respective arithmetic means of the samples X and Y, X is the original sample and Y the symmetrically modified sample according to Section 1.4.3.1.

Theorem 2

Under the null hypothesis of the Shapiro-Wilk test that the original data X comes from a normal distribution, it holds

$$\left|W_n - \tilde{W}_n\right| \xrightarrow{p} 0\,.$$

To prove Theorem 2, we prove three useful theorems.

Theorem 3

Let $\{k_n\}$ be a sequence of integers so that X_{nk_n} is a sequence of order statistics of central, intermediate or extreme terms (for definition see 1.5.3.1). Under the null hypothesis of normality, it holds

$$X_{nk_n} - Y_{nk_n} \xrightarrow{p} 0\,.$$

Theorem 4

Under the null hypothesis of normality, the sample variance of the modified sample Y converges in probability to the variance of the original sample X, i.e.

$$\frac{1}{n}\sum_{k=1}^{n} \left(Y_{nk} - \bar{Y}\right)^2 \xrightarrow{p} \sigma^2\,.$$

Theorem 5

Under the null hypothesis of normality, the mean squared difference of the order statistics of the modified sample Y and the original sample X converges in probability to 0, i.e.

$$\frac{1}{n}\sum_{k=1}^{n}(Y_{nk}-X_{nk})^2 \xrightarrow{p} 0\,.$$

1.5.3.1 Definition and auxiliary results to prove Theorem 3

Let $V=(V_1,V_2,\ldots,V_n)^t$ and $\tilde{V}=(\tilde{V}_1,\tilde{V}_2,\ldots,\tilde{V}_n)^t$ be two independent samples of length n from the uniform (0,1) distribution.

Further let V_{nk} and $\tilde{V}_{nk}$ be the k-th order statistic of V and $\tilde{V}$ respectively. Let $\{k_n\}$ be a sequence of positive integers such that $1 \le k_n \le n \quad \forall n \in \mathbb{N}$. The ratio $\frac{k_n}{n}$ is called the rank of the order statistic V_{nk_n}. If

$$p=\lim_{n\to\infty}\frac{k_n}{n}$$

exists, then p is the limiting rank of the sequence V_{nk_n}.

Let us distinguish three types of sequences of order statistics:

- sequences of central terms, with $0<p<1$
- sequences of intermediate terms, with $p=0$ and $k_n\to\infty$ or $p=1$ and $n-k_n\to\infty$
- sequences of extreme terms, if $p=0$ and k_n is bounded or $p=1$ and $n-k_n$ is bounded.

Lemma 6

If (V_{nk_n}) *and* $\left(\tilde{V}_{nk_n}\right)$ *are sequences of order statistics of the uniform distribution (0,1) of one of the three types, then*

$$n^{1/2}\left|V_{nk_n}-\tilde{V}_{nk_n}\right| \xrightarrow{p} 0\,.$$

Proof

First we consider order statistics of central terms.

Applying a corollary of Serfling (1980, page. 94) for sequences of central terms such that k_n satisfies the condition

$$\frac{k_n}{n} = p + \frac{k}{n^{1/2}} + o(\frac{1}{n^{1/2}}), \quad n \to \infty, \tag{1.19}$$

we have

$$n^{1/2}\left(V_{nk_n} - \hat{\xi}_{pn}\right) \xrightarrow{wp1} k \tag{1.20}$$

and

$$n^{1/2}\left(\tilde{V}_{nk_n} - \hat{\hat{\xi}}_{pn}\right) \xrightarrow{wp1} k,$$

where $\hat{\xi}_{pn}$ and $\hat{\hat{\xi}}_{pn}$ denote the respective sample p-th quantile of V and $\tilde{V}$.
Note that $\hat{\xi}_{pn} - \hat{\hat{\xi}}_{pn} \xrightarrow{wp1} 0$ because V and $\tilde{V}$ follow the same distribution. It follows

$$n^{1/2}\left|V_{nk_n} - \tilde{V}_{nk_n}\right| \xrightarrow{wp1} 0\,,$$

so that the result applies for sequences of central terms.
Now we consider sequences of intermediate and extreme terms.
Applying the formula for the probability density function of order statistics given in David & Nagaraja (2003), we have that the probability density function $f_{nk}(x)$ of V_{nk} is given by

$$f_{nk}(x) = \frac{n!}{(k+1)!(n-k)!}x^{k-1}(1-x)^{n-k}.$$

Thus V_{nk} is beta distributed with parameters k and $n-k+1$.
Hence from Johnson et al. (1995), we have

$$\begin{aligned} Var\left(n^{1/2}\left(V_{nk_n} - \tilde{V}_{nk_n}\right)\right) &= 2nVar\left(V_{nk_n}\right) = 2n\frac{k_n(n-k_n+1)}{(n+1)^2(n+2)} \\ &= \left(\frac{2n}{n+2}\right)\left(\frac{n}{n+1}\right)^2\left(\frac{k_n}{n}\right)\left(\frac{n-k_n+1}{n}\right) \to 0\,. \end{aligned}$$

Due to the fact that

$$E\left(n^{1/2}\left(V_{nk_n} - \tilde{V}_{nk_n}\right)\right) = n^{1/2}\left(E\left(V_{nk_n}\right) - E\left(\tilde{V}_{nk_n}\right)\right) = 0,$$

we obtain

$$\begin{aligned} & E\left[\left(n^{1/2}\left(V_{nk_n}-\tilde{V}_{nk_n}\right)\right)^2\right] = Var\left(n^{1/2}\left(V_{nk_n}-\tilde{V}_{nk_n}\right)\right) \to 0 \\ \Leftrightarrow \quad & n^{1/2}\left(V_{nk_n}-\tilde{V}_{nk_n}\right) \xrightarrow{L_2} 0 \\ \Rightarrow \quad & n^{1/2}\left(V_{nk_n}-\tilde{V}_{nk_n}\right) \xrightarrow{p} 0\,. \end{aligned}$$

Here the notation $\xrightarrow{L_2}$ means convergence in $2nd$ mean. ▮

Lemma 7

If (V_{nk_n}) and $\left(\tilde{V}_{nk_n}\right)$ are two independent sequences of order statistics of the uniform distribution (0,1) of central, intermediate or upper extreme terms, then

$$\frac{V_{nk_n}}{\tilde{V}_{nk_n}} \xrightarrow{p} 1\,. \tag{1.21}$$

Proof

For sequences of central terms so that $p = \lim_{n\to\infty}\frac{k_n}{n}$, we use a corollary of Serfling (1980, page. 94) and obtain $V_{nk_n} \xrightarrow{p} p$ and $\tilde{V}_{nk_n} \xrightarrow{p} p$. This implies the result for sequences of central terms.

For a sequence (V_{nk_n}) of lower intermediate or lower extreme terms, the Markov probability inequality yields

$$P\left(|V_{nk_n}| < \epsilon\right) \geq 1 - \frac{k_n}{(n+1)\epsilon} \to 1 \Rightarrow V_{nk_n} \xrightarrow{p} 0\,.$$

Now assume (V_{nk_n}) and $\left(\tilde{V}_{nk_n}\right)$ are sequences of upper intermediate or upper extreme terms, then $(1-V_{nk_n})$ and $\left(1-\tilde{V}_{nk_n}\right)$ are sequences of lower intermediate and lower extreme terms so that $V_{nk_n} \xrightarrow{p} 1$ and $\tilde{V}_{nk_n} \xrightarrow{p} 1$. This yields the result for sequences of upper intermediate and upper extreme terms.

It remains to show the result for (V_{nk_n}) being a sequence of lower intermediate terms. We have shown in the proof of Lemma 6 that $\tilde{V}_{nk_n}$ is beta distributed with parameters k_n and $n-k_n+1$. From Johnson et al. (1995), we have

$$\begin{aligned} E\left(\frac{1}{\tilde{V}_{nk_n}}\right) &= \frac{n}{k_n-1}, \quad k_n > 1 \\ Var\left(\frac{1}{\tilde{V}_{nk_n}}\right) &= \frac{n(n-k_n+1)}{(k_n-2)(k_n-1)^2}, \quad k_n > 2\,. \end{aligned}$$

Because V_{nk_n} and $\tilde{V}_{nk_n}$ are independent random variables, we have

$$E\left(\frac{V_{nk_n}}{\tilde{V}_{nk_n}}\right) = E\left(V_{nk_n}\right) E\left(\frac{1}{\tilde{V}_{nk_n}}\right) = \left(\frac{k_n}{n+1}\right)\left(\frac{n}{k_n-1}\right) = \left(\frac{k_n}{k_n-1}\right)\left(\frac{n}{n+1}\right) \longrightarrow 1\,.$$

From Frishman (1975), we have

$$Var\left(\frac{V_{nk_n}}{\tilde{V}_{nk_n}}\right) = [E(V_{nk_n})]^2\, Var\left(\frac{1}{\tilde{V}_{nk_n}}\right) + \left[E\left(\frac{1}{\tilde{V}_{nk_n}}\right)\right]^2 Var(V_{nk_n}) + Var(V_{nk_n})Var\left(\frac{1}{\tilde{V}_{nk_n}}\right)\,. \tag{1.22}$$

For the terms on the right hand side of equation (1.22), we obtain

$$[E(V_{nk_n})]^2\, Var\left(\frac{1}{\tilde{V}_{nk_n}}\right) = \left(\frac{k_n}{n+1}\right)^2 \frac{n(n-k_n+1)}{(k_n-2)(k_n-1)^2} = \left(\frac{k_n}{k_n-1}\right)^2\left(\frac{1}{k_n-2}\right)\frac{n(n-k_n+1)}{(n+1)^2} \to 0\,.$$

$$\left[E\left(\frac{1}{\tilde{V}_{nk_n}}\right)\right]^2 Var(V_{nk_n}) = \left(\frac{n}{k_n-1}\right)^2 \frac{k_n(n-k_n+1)}{(n+1)^2(n+2)} = \left(\frac{n}{n+1}\right)^2 \frac{k_n}{(k_n-1)^2}\left(\frac{n-k_n+1}{n+2}\right) \to 0\,.$$

$$Var(V_{nk_n})Var\left(\frac{1}{\tilde{V}_{nk_n}}\right) = \left(\frac{n-k_n+1}{n+1}\right)^2\left(\frac{n}{n+2}\right)\left(\frac{k_n}{k_n-2}\right)\frac{1}{(k_n-1)^2} \to 0\,.$$

Hence equation (1.22) yields

$$Var\left(\frac{V_{nk_n}}{\tilde{V}_{nk_n}}\right) \longrightarrow 0$$

and Chebyshev's inequality yields

$$\frac{V_{nk_n}}{\tilde{V}_{nk_n}} \xrightarrow{p} 1\,. \quad \blacksquare$$

Lemma 8

If (V_{nk_n}) and $\left(\tilde{V}_{nk_n}\right)$ are independent sequences of order statistics of the uniform distribution (0,1) of lower extreme terms ($k_n \neq 1$), then $\frac{V_{nk_n}}{\tilde{V}_{nk_n}}$ is bounded in probability, i.e

$$\forall \epsilon > 0 \quad \exists M_\epsilon \in \mathbb{R} \quad \exists N_\epsilon \in \mathbb{N} \quad \textit{so that} \quad P\left(\frac{V_{nk_n}}{\tilde{V}_{nk_n}} > M_\epsilon\right) < \epsilon \quad \forall n > N_\epsilon\,.$$

Proof

Markov's inequality yields for every fixed $\epsilon > 0$

$$P\left(\frac{V_{nk_n}}{\tilde{V}_{nk_n}} > \frac{2}{\epsilon}\right) \leq \frac{k_n}{k_n-1}\left(\frac{n}{n+1}\right)\frac{\epsilon}{2} < \left(\frac{k_n}{k_n-1}\right)\frac{\epsilon}{2} \leq \epsilon \quad \forall k_n \geq 2\,. \quad \blacksquare$$

Lemma 9

Let Φ^{-1} denote the probit function. Then

$$\lim_{x\to 0} x e^{\frac{1}{2}\left(\Phi^{-1}(x)\right)^2} = 0$$

and

$$\lim_{x\to 0} x\Phi^{-1}(x) e^{\frac{1}{2}\left(\Phi^{-1}(x)\right)^2} = -\frac{1}{\sqrt{2\pi}}\,.$$

Proof

For $0 < x < 1$ set $u = \Phi^{-1}(x)$, so that

$$\lim_{x\to 0} x e^{\frac{1}{2}\left(\Phi^{-1}(x)\right)^2} = \lim_{u\to -\infty} \Phi(u) e^{\frac{1}{2}u^2} = \lim_{u\to -\infty} \frac{\Phi(u)}{e^{-\frac{1}{2}u^2}}\,.$$

Applying the L'Hospital's rule yields

$$\lim_{u\to -\infty} \frac{\Phi(u)}{e^{-\frac{1}{2}u^2}} = \lim_{u\to -\infty} \frac{\frac{d}{du}\left(\Phi(u)\right)}{\frac{d}{du}\left(e^{-\frac{1}{2}u^2}\right)} = \lim_{u\to -\infty} \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}}{-u e^{-\frac{1}{2}u^2}} = \lim_{u\to -\infty} -\frac{1}{u\sqrt{2\pi}} = 0$$

$$\Rightarrow \quad \lim_{x\to 0} x e^{\frac{1}{2}\left(\Phi^{-1}(x)\right)^2} = 0\,.$$

In the same manner

$$\begin{aligned} \lim_{x\to 0} x\Phi^{-1}(x) e^{\frac{1}{2}\left(\Phi^{-1}(x)\right)^2} &= \lim_{u\to -\infty} u\Phi(u) e^{\frac{1}{2}u^2} = \lim_{u\to -\infty} \frac{\Phi(u)}{\frac{1}{u}e^{-\frac{1}{2}u^2}} = \lim_{u\to -\infty} \frac{\frac{d}{du}\left(\Phi(u)\right)}{\frac{d}{du}\left(\frac{1}{u}e^{-\frac{1}{2}u^2}\right)} \\ &= \lim_{u\to -\infty} \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}}{-\frac{1}{u^2}e^{-\frac{1}{2}u^2} - e^{-\frac{1}{2}u^2}} = \lim_{u\to -\infty} \frac{1}{\sqrt{2\pi}} \frac{1}{-\frac{1}{u^2} - 1} \\ \Rightarrow \lim_{x\to 0} x\Phi^{-1}(x) e^{\frac{1}{2}\left(\Phi^{-1}(x)\right)^2} &= -\frac{1}{\sqrt{2\pi}}\,. \quad \blacksquare \end{aligned}$$

Lemma 10

Assume the same conditions as in Lemma 6 and let Φ^{-1} be the probit function. Then the following holds

$$\Phi^{-1}\left(V_{nk_n}\right) - \Phi^{-1}\left(\tilde{V}_{nk_n}\right) \xrightarrow{p} 0\,. \tag{1.23}$$

Proof

Serfling (1980, page. 91) asserts

$$\Phi^{-1}\left(V_{n1}\right) + (2logn)^{1/2} \xrightarrow{wp1} 0\,.$$

Since we can write the same for $\tilde{V}_{nk_n}$, we obtain

$$\Phi^{-1}\left(V_{n1}\right) - \Phi^{-1}\left(\tilde{V}_{n1}\right) \xrightarrow{wp1} 0\,.$$

This implies the result for $k_n = 1$. For the rest of this proof, we assume $k_n > 1$.

Let $h_1(x) = \Phi^{-1}(x)$. The first and second derivatives are given by

$$\begin{aligned} h_1'(x) &= \sqrt{2\pi} e^{\frac{1}{2}\left(\Phi^{-1}(x)\right)^2} \\ h_1''(x) &= 2\pi \Phi^{-1}(x) e^{\left(\Phi^{-1}(x)\right)^2}\,. \end{aligned}$$

Further let $(x, y) \in (0,1)^2$. Taylor's theorem yields the following for the probit function

$$\Phi^{-1}(x) - \Phi^{-1}(y) = \sqrt{2\pi} e^{\frac{1}{2}\left(\Phi^{-1}(y)\right)^2}(x-y) + \frac{h_1''(\xi)}{2}(x-y)^2, \tag{1.24}$$

where ξ lies between x and y.

For $x = V_{nk_n}$ and $y = \tilde{V}_{nk_n}$ and each fixed n, we thus have

$$\Phi^{-1}\left(V_{nk_n}\right) - \Phi^{-1}\left(\tilde{V}_{nk_n}\right) = \sqrt{2\pi} e^{\frac{1}{2}\left(\Phi^{-1}(\tilde{V}_{nk_n})\right)^2}\left(V_{nk_n} - \tilde{V}_{nk_n}\right) + \frac{h_1''(\xi_n)}{2}\left(V_{nk_n} - \tilde{V}_{nk_n}\right)^2,$$

with ξ_n between V_{nk_n} and $\tilde{V}_{nk_n}$.

For sequences of central terms, equation (1.20) implies that with probability one V_{nk_n} will neither tend to 0 nor 1. It follows that with probability one $\Phi^{-1}\left(V_{nk_n}\right)$ and $\frac{h_1''(\xi_n)}{2}$ will not tend to infinity, hence using Lemma 6

$$\Phi^{-1}\left(V_{nk_n}\right) - \Phi^{-1}\left(\tilde{V}_{nk_n}\right) \xrightarrow{wp1} 0,$$

which implies equation (1.23) for sequences of central terms.

Due to the symmetry of the uniform distribution around its mean and the symmetry of the probit function around 1/2, we can conduct the proof for sequences of lower intermediate and extreme terms and the result for the upper intermediate and extreme terms will follow by symmetry. Let us consider V_{nk_n} to be a sequence of lower intermediate or extreme terms. We know that $V_{nk_n} \xrightarrow{p} 0$.

Define

$$h_2(x) = \begin{cases} xe^{\frac{1}{2}\left(\Phi^{-1}(x)\right)^2}, & \text{if } 0 < x < 1 \\ 0, & \text{if } x = 0 \end{cases}$$

Lemma 9 implies that $h_2(x)$ is a continuous function and the continuous mapping theorem assures

$$h_2(V_{nk_n}) \xrightarrow{p} h_2(0) \Leftrightarrow V_{nk_n} e^{\frac{1}{2}\left(\Phi^{-1}(V_{nk_n})\right)^2} \xrightarrow{p} 0\,.$$

Because of Lemma 7 and Lemma 8, we know that either $\frac{\tilde{V}_{nk_n}}{V_{nk_n}}$ is bounded in probability or $\frac{\tilde{V}_{nk_n}}{V_{nk_n}} \xrightarrow{p} 1$. In both cases, we obtain

$$\frac{\tilde{V}_{nk_n}}{V_{nk_n}} V_{nk_n} e^{\frac{1}{2}\left(\Phi^{-1}(V_{nk_n})\right)^2} \xrightarrow{p} 0 \Rightarrow \tilde{V}_{nk_n} e^{\frac{1}{2}\left(\Phi^{-1}(V_{nk_n})\right)^2} \xrightarrow{p} 0,$$

Hence

$$\sqrt{2\pi} e^{\frac{1}{2}\left(\Phi^{-1}(\tilde{V}_{nk_n})\right)^2} \left(V_{nk_n} - \tilde{V}_{nk_n}\right) \xrightarrow{p} 0\,. \tag{1.25}$$

To finish the proof, we just need to show that

$$\frac{h_1''(\xi_n)}{2} \left(V_{nk_n} - \tilde{V}_{nk_n}\right)^2 \xrightarrow{p} 0\,. \tag{1.26}$$

holds for order statistics of intermediate and extreme terms.

It is only necessary to show

$$\frac{h_1''(\xi_n)}{2} \left(V_{nk_n}\right)^2 \xrightarrow{p} 0,$$

because due to Lemma 7 and Lemma 8, it will immediately follow

$$\begin{gathered} \frac{h_1''(\xi_n)}{2} \left(\tilde{V}_{nk_n}\right)^2 \xrightarrow{p} 0 \\ \frac{h_1''(\xi_n)}{2} \left(V_{nk_n} \tilde{V}_{nk_n}\right) \xrightarrow{p} 0\,. \end{gathered}$$

Define

$$h_3(x) = \begin{cases} x^2\Phi^{-1}(x)e^{\left(\Phi^{-1}(x)\right)^2}, & \text{if } 0 < x < 1 \\ 0, & \text{if } x = 0 \end{cases}$$

Lemma 9 implies that

$$\lim_{x\to 0} h_3(x) = \lim_{x\to 0} x^2\Phi^{-1}(x)e^{\left(\Phi^{-1}(x)\right)^2} = \left(\lim_{x\to 0} xe^{\frac{1}{2}\left(\Phi^{-1}(x)\right)^2}\right)\left(\lim_{x\to 0} x\Phi^{-1}(x)e^{\frac{1}{2}\left(\Phi^{-1}(x)\right)^2}\right) = 0\,.$$

This means that $h_3(x)$ is a continuous function and again with the continuous mapping theorem, we have

$$V_{nk_n} \xrightarrow{p} 0 \Rightarrow h_3(V_{nk_n}) \xrightarrow{p} h_3(0) \Leftrightarrow \frac{h_1''(V_{nk_n})}{2}\,(V_{nk_n})^2 \xrightarrow{p} 0\,.$$

Lemma 7 and Lemma 8 yield the equivalent result

$$\frac{h_1''(\tilde{V}_{nk_n})}{2}\,(V_{nk_n})^2 \xrightarrow{p} 0\,.$$

Since ξ_n lies between V_{nk_n} and $\tilde{V}_{nk_n}$ and it is easy to show that h_1'' is a strictly increasing function, we obtain

$$\frac{h_1''(\xi_n)}{2}\,(V_{nk_n})^2 \xrightarrow{p} 0\,. \quad \blacksquare$$

Lemma 11

Recalling the notations of Section 1.5, we state

$$X_{nk_n} - \tilde{X}_{nk_n} \xrightarrow{p} 0\,. \tag{1.27}$$

Proof

Let $V_i = \Phi\left(\frac{X_i-\mu}{\sigma}\right)$ and $\tilde{V}_i = \Phi\left(\frac{\tilde{X}_i-\hat{\mu}_n}{\hat{\sigma}_n}\right)$. Then $V_1,\ldots,V_n$ and $\tilde{V}_1,\ldots,\tilde{V}_n$ both are independent uniform(0,1) samples and because of monotonicity

$$\begin{aligned} X_{nk_n} &= F^{-1}\left(V_{nk_n}\right) = \mu + \sigma\Phi^{-1}(V_{nk_n}) \\ \tilde{X}_{nk_n} &= G_n^{-1}\left(\tilde{V}_{nk_n}\right) = \hat{\mu}_n + \hat{\sigma}_n\Phi^{-1}(\tilde{V}_{nk_n})\,. \end{aligned}$$

Using these equalities, we obtain

$$\begin{aligned} X_{nk_n} - \tilde{X}_{nk_n} &= \mu + \sigma\Phi^{-1}(V_{nk_n}) - \left(\hat{\mu}_n + \hat{\sigma}_n\Phi^{-1}(\tilde{V}_{nk_n})\right) \\ &= [\mu - \hat{\mu}_n] + \left[\sigma\left(\Phi^{-1}\left(V_{nk_n}\right) - \Phi^{-1}\left(\tilde{V}_{nk_n}\right)\right)\right] + \left[(\sigma - \hat{\sigma}_n)\Phi^{-1}(\tilde{V}_{nk_n})\right] . \end{aligned}$$

As a consequence of Lemma 10 we have

$$\sigma\left(\Phi^{-1}\left(V_{nk_n}\right) - \Phi^{-1}\left(\tilde{V}_{nk_n}\right)\right) \xrightarrow{p} 0 .$$

Serfling & Mazumder (2009) showed that $\mu - \hat{\mu}_n \xrightarrow{wp1} 0$ and $\sigma - \hat{\sigma}_n \xrightarrow{wp1} 0$ at an exponential rate. We also know from (Serfling, 1980, page. 91) that

$$\Phi^{-1}\left(V_{nn}\right) - \left(2\log(n)\right)^{1/2} \xrightarrow{wp1} 0,$$

so that it follows

$$(\sigma - \hat{\sigma}_n)\Phi^{-1}(\tilde{V}_{nn}) \xrightarrow{p} 0. \tag{1.28}$$

Given that $\Phi^{-1}\left(V_{nn}\right)$ and $\Phi^{-1}\left(V_{n1}\right)$ have the fastest rate of convergence to infinity (the convergence rate being the same for both quantities due to symmetry), equation (1.28) implies

$$(\sigma - \hat{\sigma}_n)\Phi^{-1}(\tilde{V}_{nk_n}) \xrightarrow{p} 0$$

and the result follows

$$X_{nk_n} - \tilde{X}_{nk_n} \xrightarrow{p} 0 . \quad \blacksquare$$

1.5.3.2 Proof of theorem 3

Proof

The modified sequence Y is composed of three ordered parts

- Part 1: $\tilde{X}_{n,1} \leq \tilde{X}_{n,2} \leq \cdots \leq \tilde{X}_{n,L_n}$

- Part 2: $X_{n,L_n+1} \leq \cdots \leq X_{n,n-U_n}$

- Part 3: $\tilde{X}_{n,n-U_n+1} \leq \tilde{X}_{n,n-U_n+2} \leq \cdots \leq \tilde{X}_{n,n}$

Note that the modified sequence is not ordered yet. Also notice that $\tilde{X}_{n,L_n}$ and X_{n,L_n+1} are order statistics of central terms because Lemma 4 states that

$$\frac{L_n}{n} \xrightarrow{wp1} \Phi(-3)\,.$$

Let us focus on ordering the two first parts of the modified sequence.
To do this, we search for an upper bound for $\tilde{X}_{n,L_n}$ in Part 2 and a lower bound for X_{n,L_n+1} in Part 1.
Let us first consider the case of $\tilde{X}_{n,L_n}$.
Let $\{c_n\}$ be a positive sequence of integers such that $L_n \leq c_n \leq n-U_n$ and c be a strictly positive real number. We show that if $\lim_{n\to\infty} X_{n,c_n}$ is not larger than $\lim_{n\to\infty} \tilde{X}_{n,L_n}$ with probability one then

$$\forall c > 0: \quad \frac{c_n}{n} < \Phi(-3) + \frac{c}{n^{1/2}} + o(\frac{1}{n^{1/2}}), \quad n \to \infty$$

and

$$n^{1/2}\left(X_{n,c_n} - \tilde{X}_{n,L_n}\right) \xrightarrow{wp1} 0\,.$$

Suppose $\{c_n\}$ is a positive sequence of integers such that $L_n \leq c_n \leq n-U_n$ and fulfilling

$$\exists c > 0: \quad \frac{c_n}{n} \geq \Phi(-3) + \frac{c}{n^{1/2}} + o(\frac{1}{n^{1/2}}), \quad n \to \infty\,. \tag{1.29}$$

Let us consider the case of equality in equation (1.29). Then we apply a corollary of Serfling (1980, page 94) to obtain

$$n^{1/2}\left(X_{n,c_n} - (\mu - 3\sigma)\right) \xrightarrow{wp1} \frac{c\sigma}{\phi(-3)}$$

and

$$n^{1/2}\left(\tilde{X}_{n,L_n} - (\mu - 3\sigma)\right) \xrightarrow{wp1} 0,$$

where $\phi(x)$ is the density function of the standard normal distribution.

$$\Rightarrow \lim_{n\to\infty} \quad n^{1/2}\left(X_{n,c_n} - (\mu - 3\sigma)\right) > \lim_{n\to\infty} n^{1/2}\left(\tilde{X}_{n,L_n} - (\mu - 3\sigma)\right) \quad \text{with probability one}$$

$$\Leftrightarrow \quad \lim_{n\to\infty} X_{n,c_n} > \lim_{n\to\infty} \tilde{X}_{n,L_n} \quad \text{with probability one.}$$

It is obvious that this result will still hold if strict inequality is fulfilled in equation (1.29). Thus we have shown the equivalent result that if X_{n,c_n} is not larger than $\tilde{X}_{n,L_n}$ with probability one then the constant c must be null and

$$\frac{L_n}{n} \leq \frac{c_n}{n} = \Phi(-3) + o(\frac{1}{n^{1/2}}), \quad n \to \infty \tag{1.30}$$

and hence

$$n^{1/2}\left(X_{n,c_n} - \tilde{X}_{n,L_n}\right) \xrightarrow{wp1} 0 \,.$$

By analogy, for the lower bound of X_{n,L_n+1}, we have that for a sequence of integers $\{c'_n\}$ such that $1 \leq c'_n \leq L_n$, if $\lim_{n\to\infty} \tilde{X}_{n,c'_n}$ is not smaller than $\lim_{n\to\infty} X_{n,L_n+1}$ with probability one, then

$$\frac{L_n}{n} \geq \frac{c'_n}{n} = \Phi(-3) + o(\frac{1}{n^{1/2}}), \quad n \to \infty \tag{1.31}$$

and it would follow

$$n^{1/2}\left(X_{n,L_n+1} - \tilde{X}_{n,c'_n}\right) \xrightarrow{wp1} 0 \,.$$

So that arranging the first two parts of the modified sequence in ascending order as n grows to infinity can be done in the following manner

- $Y_{n,k_n} = \tilde{X}_{n,k_n}$ for $1 \leq k_n \leq c'_n - 1$, the set of observations of $\lim_{n\to\infty} \tilde{X}$ that are smaller than $\lim_{n\to\infty} X_{n,L_n+1}$ with probability one and it follows from Lemma 11

$$Y_{n,k_n} - X_{n,k_n} = \tilde{X}_{n,k_n} - X_{n,k_n} \xrightarrow{p} 0 \,.$$

- $Y_{n,k_n} = X_{n,k_n}$ for $c_n + 1 \leq k_n \leq n - U_n$, the set of observations of $\lim_{n\to\infty} X$ that are larger than $\lim_{n\to\infty} \tilde{X}_{n,L_n}$ with probability one, so that

$$Y_{n,k_n} - X_{n,k_n} = X_{n,k_n} - X_{n,k_n} = 0 \xrightarrow{p} 0 \,.$$

- $Y_{n,k_n} \in \{\tilde{X}_{n,c'_n}, \ldots, \tilde{X}_{n,L_n}, X_{n,L_n+1}, \ldots, X_{n,c_n}\}$ such that equations (1.30) and (1.31) hold and hence, we conclude

$$n^{1/2}\left(Y_{n,k_n} - X_{n,k_n}\right) \xrightarrow{wp1} 0 \,.$$

We can apply the same logic to arrange Part 2 and 3 of the modified sequence to have that

$$Y_{n,k_n} - X_{n,k_n} \xrightarrow{p} 0$$

for any sequence of order statistics of one of the three types. ▮

1.5.3.3 Auxiliary results to prove Theorem 4

Lemma 12

Let $T_{n1}, T_{n2}, \ldots, T_{nn}$ and $\tilde{T}_{n1}, \tilde{T}_{n2}, \ldots, \tilde{T}_{nn}$ be the order statistics of independent samples of independent and identically standard normal distributed random variables. Define T_{nk} and $\tilde{T}_{nk}$ as the k-th order statistics in independent samples of length n. We have

$$\frac{1}{n}\sum_{k=1}^{n} T_{nk}\tilde{T}_{nk} \xrightarrow{p} 1\,.$$

Proof

Let us define

$$\Psi_n = \frac{1}{n}\sum_{k=1}^{n} T_{nk}\tilde{T}_{nk}\,.$$

We will apply Chebyshev's inequality to Ψ_n after computing its expectation and showing that

$$\lim_{n\to\infty} Var(\Psi_n) = 0\,.$$

We have

$$\begin{aligned}
E(\Psi_n) &= \frac{1}{n}\sum_{k=1}^{n} E\left(T_{nk}\right) E\left(\tilde{T}_{nk}\right) = \frac{1}{n}\sum_{k=1}^{n} \left[E\left(T_{nk}\right)\right]^2 = \frac{1}{n}\sum_{k=1}^{n} \left[E\left(T_{nk}^2\right) - Var(T_{nk})\right] \\
&= \frac{1}{n}E\left(\sum_{k=1}^{n} T_{nk}^2\right) - \frac{1}{n}\sum_{k=1}^{n} Var(T_{nk}) = \frac{1}{n}E\left(\sum_{k=1}^{n} T_{k}^2\right) - \frac{1}{n}\sum_{k=1}^{n} Var(T_{nk}) \\
E(\Psi_n) &= 1 - \frac{1}{n}\sum_{k=1}^{n} Var(T_{nk})\,. \qquad (1.32)
\end{aligned}$$

Stephens (1975) shows that the asymptotic eigenvalues of the covariance matrix V_0 of the order statistics are values of the sequence $\{\frac{1}{n}\}_{n\in\mathbb{N}}$ so that its trace fulfils

$$\frac{1}{n}\sum_{k=1}^{n} Var(T_{nk}) - \frac{1}{n}\sum_{k=1}^{n}\frac{1}{k} \xrightarrow{n\to\infty} 0\,. \tag{1.33}$$

The second sum in (1.33) is a Cesaro mean (see Cesàro (1888)) and it follows

$$\begin{aligned}\frac{1}{k} \xrightarrow{k\to\infty} 0 \Rightarrow \frac{1}{n}\sum_{k=1}^{n}\frac{1}{k} \xrightarrow{n\to\infty} 0 \\ \Rightarrow \frac{1}{n}\sum_{k=1}^{n} Var(T_{nk}) \xrightarrow{n\to\infty} 0\,. \end{aligned} \tag{1.34}$$

Hence from equations (1.32) and (1.34), we can conclude that Ψ_n has a finite mean and it holds

$$E(\Psi_n) = 1 - \frac{1}{n}\sum_{k=1}^{n} Var(T_{nk}) \xrightarrow{n\to\infty} 1\,.$$

We know that

$$Var(\Psi_n) = \frac{1}{n^2}\sum_{k=1}^{n} Var(T_{nk}\tilde{T}_{nk}) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} Cov\left(T_{ni}\tilde{T}_{ni}, T_{nj}\tilde{T}_{nj}\right)\,. \tag{1.35}$$

And from Frishman (1975), we have

$$\begin{aligned} Var(T_{nk}\tilde{T}_{nk}) &= [E(T_{nk})]^2\, Var(\tilde{T}_{nk}) + \left[E(\tilde{T}_{nk})\right]^2 Var(T_{nk}) + Var(T_{nk})Var(\tilde{T}_{nk}) \\ &= 2\,[E(T_{nk})]^2\, Var(T_{nk}) + (Var(T_{nk}))^2 \\ &= 2\left[E(T_{nk}^2) - Var(T_{nk})\right] Var(T_{nk}) + (Var(T_{nk}))^2 \\ Var(T_{nk}\tilde{T}_{nk}) &= 2E(T_{nk}^2)Var(T_{nk}) - (Var(T_{nk}))^2\,. \end{aligned} \tag{1.36}$$

We have

$$\frac{1}{n^2}\sum_{k=1}^{n}(Var(T_{nk}))^2 = \left[\frac{1}{n}\left(\sum_{k=1}^{n}(Var(T_{nk}))^2\right)^{1/2}\right]^2 \le \left[\frac{1}{n}\sum_{k=1}^{n} Var(T_{nk})\right]^2 \xrightarrow{n\to\infty} 0 \tag{1.37}$$

$$\frac{1}{n^2}\sum_{k=1}^{n} E(T_{nk}^2)Var(T_{nk}) \le \frac{1}{n^2}\sum_{k=1}^{n} E(T_{nk}^2)Var(T_{nk}) + \frac{1}{n^2}\sum_{k=1}^{n}\left(E(T_{nk}^2)\left[\sum_{\substack{j=1\\j\neq k}}^{n} Var(T_{nj})\right]\right)$$

$$= \left(\frac{1}{n}\sum_{k=1}^{n} E(T_{nk}^2)\right)\left(\frac{1}{n}\sum_{k=1}^{n} Var(T_{nk})\right) = \frac{1}{n}\sum_{k=1}^{n} Var(T_{nk})$$

$$\Rightarrow \frac{1}{n^2}\sum_{k=1}^{n} E(T_{nk}^2)Var(T_{nk}) \le \frac{1}{n}\sum_{k=1}^{n} Var(T_{nk})$$

with equation (1.34), we obtain

$$0 \le \frac{1}{n^2}\sum_{k=1}^{n} E(T_{nk}^2)Var(T_{nk}) \xrightarrow{n\to\infty} 0\,. \tag{1.38}$$

Equations (1.36), (1.37) and (1.38) imply

$$\frac{1}{n^2}\sum_{k=1}^{n} Var(T_{nk}\tilde{T}_{nk}) \xrightarrow{n\to\infty} 0\,. \tag{1.39}$$

We now consider the second term on right hand side of equation (1.35). It holds

$$\begin{aligned}
Cov\left(T_{ni}\tilde{T}_{ni}, T_{nj}\tilde{T}_{nj}\right) &= E\left(T_{ni}\tilde{T}_{ni}T_{nj}\tilde{T}_{nj}\right) - E\left(T_{ni}\tilde{T}_{ni}\right)E\left(T_{nj}\tilde{T}_{nj}\right)\\
&= \left[E\left(T_{ni}T_{nj}\right)\right]^2 - \left[E\left(T_{ni}\right)\right]^2\left[E\left(T_{nj}\right)\right]^2\\
&= \left[Cov\left(T_{ni}, T_{nj}\right) + E(T_{ni})E(T_{nj})\right]^2 - \left[E\left(T_{ni}\right)\right]^2\left[E\left(T_{nj}\right)\right]^2\\
Cov\left(T_{ni}\tilde{T}_{ni}, T_{nj}\tilde{T}_{nj}\right) &= \left[Cov\left(T_{ni}, T_{nj}\right)\right]^2 + 2Cov\left(T_{ni}, T_{nj}\right)E(T_{ni})E(T_{nj})\,.
\end{aligned} \tag{1.40}$$

For the first term on the right hand side of equation (1.40) we can write

$$\begin{aligned}
\frac{1}{n^2}\sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n}\left[Cov\left(T_{ni}, T_{nj}\right)\right]^2 &\le \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[Cov\left(T_{ni}, T_{nj}\right)\right]^2\\
&\le \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} Var(T_{ni})Var(T_{nj}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(T_{ni})\sum_{i=1}^{n} Var(T_{nj})\\
&= \left(\frac{1}{n}\sum_{i=1}^{n} Var(T_{ni})\right)^2 \xrightarrow{n\to\infty} 0\,.
\end{aligned} \tag{1.41}$$

Let m denote the vector of expected values of order statistics from a standard normal distribution and recall that V_0 is the corresponding covariance matrix. For the second term on the right hand side of equation (1.40) we have

$$\sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n} Cov\,(T_{ni},T_{nj})\,E(T_{ni})E(T_{nj}) = m^tV_0m - \sum_{i=1}^{n} Cov\,(T_{ni},T_{ni})\,E(T_{ni})E(T_{ni}) \tag{1.42}$$

where

$$m^tV_0m = \sum_{i=1}^{n}\sum_{j=1}^{n} Cov\,(T_{ni},T_{nj})\,E(T_{ni})E(T_{nj})\,.$$

As shown by Stephens (1975)

$$\|V_0m - \frac{1}{2}m\| \xrightarrow{n\to\infty} 0,$$

where $\| \bullet \|$ is the Euclidean norm. Due to the fact that

$$m^tm - n \xrightarrow{n\to\infty} 0,$$

it follows

$$\frac{1}{n^2}m^tV_0m = \frac{1}{n^2}m^t\left(V_0m - \frac{1}{2}m\right) + \frac{1}{2n^2}m^tm \xrightarrow{n\to\infty} 0\,.$$

On the other hand , we use equation (1.38) to obtain

$$\frac{1}{n^2}\sum_{i=1}^{n} Cov\,(T_{ni},T_{ni})\,E(T_{ni})E(T_{ni}) = \frac{1}{n^2}\sum_{i=1}^{n} Var\,(T_{ni})\,(E(T_{ni}))^2 \leq \frac{1}{n^2}\sum_{i=1}^{n} Var\,(T_{ni})\,E(T_{ni}^2) \xrightarrow{n\to\infty} 0\,.$$

Hence equation (1.42) yields

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n} Cov\,(T_{ni},T_{nj})\,E(T_{ni})E(T_{nj}) \xrightarrow{n\to\infty} 0\,. \tag{1.43}$$

Equations (1.40), (1.41) and (1.43) lead to

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n} Cov\left(T_{ni}\tilde{T}_{ni}, T_{nj}\tilde{T}_{nj}\right) \xrightarrow{n\to\infty} 0\,. \tag{1.44}$$

Finally it follows from equations (1.35), (1.39) and (1.44) that

$$Var(\Psi_n) \xrightarrow{n\to\infty} 0\,. \tag{1.45}$$

Applying Chebyshev's inequality to Ψ_n and considering equation (1.45), we have the following

$$\forall \epsilon > 0 : \quad P(|\Psi_n - E(\Psi_n)| > \epsilon) \leq \frac{Var(\Psi_n)}{\epsilon^2} \xrightarrow{n\to\infty} 0$$

which is equivalent to

$$\Psi_n - 1 + \frac{1}{n}\sum_{k=1}^{n} Var(X_{nk}) \xrightarrow{p} 0$$

and due to equation (1.34)

$$\Psi_n \xrightarrow{p} 1 \,. \quad \blacksquare$$

Lemma 13

Given the notations of Section 1.5, firstly, the arithmetic means of the observed and the artificial samples converge weakly to the mean of the observed sample,i.e

$$\frac{1}{n}\sum_{k=1}^{n} \tilde{X}_{nk} \xrightarrow{p} \mu \tag{1.46}$$

$$\frac{1}{n}\sum_{k=1}^{n} \left(X_{nk} - \tilde{X}_{nk} \right) \xrightarrow{p} 0 \,. \tag{1.47}$$

Secondly, the sample means of squares of the observed and artificial samples converge weakly to the second order moment of the observed sample computed under the normality assumption, i.e

$$\frac{1}{n}\sum_{k=1}^{n} \tilde{X}_{nk}^2 \xrightarrow{p} \mu^2 + \sigma^2 \tag{1.48}$$

$$\frac{1}{n}\sum_{k=1}^{n} \left(X_{nk}^2 - \tilde{X}_{nk}^2 \right) \xrightarrow{p} 0 \,. \tag{1.49}$$

Proof

Because conditionally on $(\hat{\mu}_n, \hat{\sigma}_n^2)$ it holds $\frac{1}{n}\sum_{k=1}^{n} \tilde{X}_k \sim \mathcal{N}\left(\hat{\mu}_n, \frac{\hat{\sigma}_n^2}{n}\right)$, we obtain with Chebyshev's inequality

$$\frac{1}{n}\sum_{k=1}^{n} \tilde{X}_{nk} - \hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} \tilde{X}_k - \hat{\mu}_n \xrightarrow{p} 0 \,.$$

Due to the strong convergence $\hat{\mu}_n \xrightarrow{wp1} \mu$, we obtain

$$\frac{1}{n}\sum_{k=1}^{n} \tilde{X}_{nk} \xrightarrow{p} \mu$$

and equation (1.46) is proven.

The same reasoning yields

$$\frac{1}{n}\sum_{k=1}^{n} X_{nk} = \frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{p} \mu \,.$$

Hence we have

$$\frac{1}{n}\sum_{k=1}^{n}\left(X_{nk} - \tilde{X}_{nk}\right) = \frac{1}{n}\sum_{k=1}^{n} X_{nk} - \frac{1}{n}\sum_{k=1}^{n} \tilde{X}_{nk} \xrightarrow{p} 0,$$

so that equation (1.47) is proven.

After straightforward calculations, we obtain

$$E\left(\frac{1}{n}\sum_{k=1}^{n} \tilde{X}_{nk}^2\right) = E\left(\frac{1}{n}\sum_{k=1}^{n} \tilde{X}_{k}^2\right) = \hat{\mu}_n^2 + \hat{\sigma}_n^2 < \infty$$
$$Var\left(\frac{1}{n}\sum_{k=1}^{n} \tilde{X}_{nk}^2\right) = Var\left(\frac{1}{n}\sum_{k=1}^{n} \tilde{X}_{k}^2\right) = \frac{1}{n^2}\sum_{k=1}^{n} Var(\tilde{X}_k^2) = \frac{1}{n}Var(\tilde{X}^2) = \frac{2\hat{\sigma}_n^4 + 4\hat{\sigma}_n^2\hat{\mu}_n^2}{n} \xrightarrow{n\to\infty} 0 \,.$$

Then Chebyshev's inequality yields

$$\frac{1}{n}\sum_{k=1}^{n} \tilde{X}_{nk}^2 - \left(\hat{\mu}_n^2 + \hat{\sigma}_n^2\right) \xrightarrow{p} 0 \,.$$

Together with the strong convergence $\hat{\mu}_n^2 + \hat{\sigma}_n^2 \xrightarrow{wp1} \mu^2 + \sigma^2$, equation (1.48) follows.

In the same manner, we show

$$\frac{1}{n}\sum_{k=1}^{n} X_{nk}^2 \xrightarrow{p} \mu^2 + \sigma^2$$

and equation (1.49) also follows. ▮

Lemma 14

The observed data $X_{n1}, X_{n2}, \ldots, X_{nn}$ and the artificial data $\tilde{X}_{n1}, \tilde{X}_{n2}, \ldots, \tilde{X}_{nn}$ fulfil, as n goes to infinity,

$$\frac{1}{n}\sum_{k=1}^{n} X_{nk}\tilde{X}_{nk} \xrightarrow{p} \sigma^2 + \mu^2 \,.$$

Proof

From Lemma 12, we get

$$\frac{1}{n}\sum_{k=1}^{n}\left(\frac{X_{nk}-\mu}{\sigma}\right)\left(\frac{\tilde{X}_{nk}-\hat{\mu}_n}{\hat{\sigma}_n}\right)\xrightarrow{p}1 \tag{1.50}$$

because $\sigma\hat{\sigma}_n \xrightarrow{p} \sigma^2$, equation (1.50) implies

$$\begin{aligned}&\frac{1}{n}\sum_{k=1}^{n}\left(X_{nk}-\mu\right)\left(\tilde{X}_{nk}-\hat{\mu}_n\right)\xrightarrow{p}\sigma^2\\ \Leftrightarrow&\frac{1}{n}\sum_{k=1}^{n}X_{nk}\tilde{X}_{nk}-\hat{\mu}_n\frac{1}{n}\sum_{k=1}^{n}X_{nk}-\mu\frac{1}{n}\sum_{k=1}^{n}\left(\tilde{X}_{nk}-\hat{\mu}_n\right)\xrightarrow{p}\sigma^2\,.\end{aligned} \tag{1.51}$$

With Lemma 12 and the strong convergence $\hat{\mu}_n \xrightarrow{wp1} \mu$, we have

$$\begin{aligned}&\mu\frac{1}{n}\sum_{k=1}^{n}\left(\tilde{X}_{nk}-\hat{\mu}_n\right)\xrightarrow{p}0 \quad \text{and}\\ &\hat{\mu}_n\frac{1}{n}\sum_{k=1}^{n}X_{nk}\xrightarrow{p}\mu^2\end{aligned}$$

and together with equation (1.51), it follows

$$\frac{1}{n}\sum_{k=1}^{n}X_{nk}\tilde{X}_{nk}\xrightarrow{p}\sigma^2+\mu^2\,. \quad \blacksquare$$

Lemma 15

The mean squared differences between the order statistics of the observed and the artificial sample converge weakly to 0, i.e

$$\frac{1}{n}\sum_{k=1}^{n}\left(X_{nk}-\tilde{X}_{nk}\right)^2\xrightarrow{p}0\,.$$

Proof

With Lemma 14 and Lemma 13 equations (1.48) and (1.49), we can write

$$\begin{aligned}&\frac{1}{n}\sum_{k=1}^{n}\left(X_{nk}^2-\tilde{X}_{nk}^2\right)\xrightarrow{p}0\\ &\frac{1}{n}\sum_{k=1}^{n}\tilde{X}_{nk}^2-\frac{1}{n}\sum_{k=1}^{n}X_{nk}\tilde{X}_{nk}\xrightarrow{p}0\,.\end{aligned}$$

It follows

$$\frac{1}{n}\sum_{k=1}^{n}\left(X_{nk}-\tilde{X}_{nk}\right)^2=\left(\frac{1}{n}\sum_{k=1}^{n}\left(X_{nk}^2-\tilde{X}_{nk}^2\right)\right)+2\left(\frac{1}{n}\sum_{k=1}^{n}\tilde{X}_{nk}^2-\frac{1}{n}\sum_{k=1}^{n}X_{nk}\tilde{X}_{nk}\right)\xrightarrow{p}0\,. \quad \blacksquare$$

1.5.3.4 Proof of Theorem 4

Proof

First we prove the convergence of the sample mean.

$$\frac{1}{n}\sum_{k=1}^{n} Y_i = \frac{1}{n}\sum_{k=1}^{L_n} \tilde{X}_{nk} + \frac{1}{n}\sum_{k=L_n+1}^{n-U_n} X_{nk} + \frac{1}{n}\sum_{k=n-U_n+1}^{n} \tilde{X}_{nk}\,.$$

Hence

$$\begin{aligned}\left|\frac{1}{n}\sum_{k=1}^{n} Y_k - \frac{1}{n}\sum_{k=1}^{n} X_k\right| &= \left|\frac{1}{n}\sum_{k=1}^{L_n}\left(\tilde{X}_{nk} - X_{nk}\right) + \frac{1}{n}\sum_{k=n-U_n+1}^{n}\left(\tilde{X}_{nk} - X_{nk}\right)\right| \\ &\leq \frac{1}{n}\sum_{k=1}^{L_n}\left|\tilde{X}_{nk} - X_{nk}\right| + \frac{1}{n}\sum_{k=n-U_n+1}^{n}\left|\tilde{X}_{nk} - X_{nk}\right| \\ &\leq \frac{1}{n}\sum_{k=1}^{n}\left|\tilde{X}_{nk} - X_{nk}\right| \leq \left(\frac{1}{n}\sum_{k=1}^{n}\left(\tilde{X}_{nk} - X_{nk}\right)^2\right)^{1/2}\end{aligned}$$

and Lemma 15 yields

$$\left|\frac{1}{n}\sum_{k=1}^{n} Y_i - \frac{1}{n}\sum_{k=1}^{n} X_i\right| \xrightarrow{p} 0\,.$$

Joined to the fact that $\frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{wp1} \mu$, we have

$$\frac{1}{n}\sum_{k=1}^{n} Y_k \xrightarrow{p} \mu\,. \tag{1.52}$$

In the same manner, using Hölder's inequality we obtain

$$\begin{aligned}\left|\frac{1}{n}\sum_{k=1}^{n} Y_k^2 - \frac{1}{n}\sum_{k=1}^{n} X_k^2\right| &\leq \frac{1}{n}\sum_{k=1}^{L_n}\left|\tilde{X}_{nk}^2 - X_{nk}^2\right| + \frac{1}{n}\sum_{k=n-U_n+1}^{n}\left|\tilde{X}_{nk}^2 - X_{nk}^2\right| \\ &\leq \frac{1}{n}\sum_{k=1}^{n}\left|\tilde{X}_{nk}^2 - X_{nk}^2\right| = \frac{1}{n}\sum_{k=1}^{n}\left|\tilde{X}_{nk} + X_{nk}\right|\left|\tilde{X}_{nk} - X_{nk}\right|\end{aligned}$$

$$\left|\frac{1}{n}\sum_{k=1}^{n} Y_k^2 - \frac{1}{n}\sum_{k=1}^{n} X_k^2\right| \leq \left(\frac{1}{n}\sum_{k=1}^{n}\left(\tilde{X}_{nk} + X_{nk}\right)^2\right)^{1/2}\left(\frac{1}{n}\sum_{k=1}^{n}\left(\tilde{X}_{nk} - X_{nk}\right)^2\right)^{1/2} \tag{1.53}$$

$$\frac{1}{n}\sum_{k=1}^{n}\left(\tilde{X}_{nk} + X_{nk}\right)^2 = \frac{1}{n}\sum_{k=1}^{n}\tilde{X}_{nk}^2 + \frac{1}{n}\sum_{k=1}^{n} X_{nk}^2 + \frac{2}{n}\sum_{k=1}^{n}\tilde{X}_{nk}X_{nk}\,. \tag{1.54}$$

Using Lemma 14 and Lemma 13 in equation (1.54) leads us to

$$\frac{1}{n}\sum_{k=1}^{n}\left(\tilde{X}_{nk}+X_{nk}\right)^2 \xrightarrow{p} 4\left(\mu^2+\sigma^2\right) .$$

Hence due to Lemma 15 and equation (1.53), we have

$$\left|\frac{1}{n}\sum_{k=1}^{n}Y_k^2-\frac{1}{n}\sum_{k=1}^{n}X_k^2\right| \xrightarrow{p} 0 .$$

Because $\frac{1}{n}\sum_{k=1}^{n}X_k^2 \xrightarrow{wp1} \sigma^2+\mu^2$, we obtain

$$\frac{1}{n}\sum_{k=1}^{n}Y_k^2 \xrightarrow{p} \sigma^2+\mu^2 . \tag{1.55}$$

Using equations (1.52) and (1.55), it is straightforward that

$$\frac{1}{n}\sum_{k=1}^{n}\left(Y_{nk}-\bar{Y}\right)^2=\left(\frac{1}{n}\sum_{k=1}^{n}Y_k^2\right)-\left(\frac{1}{n}\sum_{k=1}^{n}Y_k\right)^2 \xrightarrow{p} \sigma^2+\mu^2-\mu^2=\sigma^2 . \quad \blacksquare$$

1.5.3.5 Proof of Theorem 5

Proof

Recalling the proof of Theorem 3, we stated that the modified sequence Y is composed of three ordered parts

- Part 1: $\tilde{X}_{n,1} \le \tilde{X}_{n,2} \le \cdots \le \tilde{X}_{n,L_n}$
- Part 2: $X_{n,L_n+1} \le \cdots \le X_{n,n-U_n}$
- Part 3: $\tilde{X}_{n,n-U_n+1} \le \tilde{X}_{n,n-U_n+2} \le \cdots \le \tilde{X}_{n,n}$

We focused on ordering the first two parts Part 1 and 2 in ascending order and we have

- $Y_{n,k}=\tilde{X}_{n,k}$ for $1 \le k \le c'_n-1$, the set of observations of $\tilde{X}$ that are smaller than $\lim_{n\to\infty} X_{n,L_n+1}$ with probability one and it follows

$$\frac{1}{n}\sum_{k=1}^{c'_n-1}\left(Y_{nk}-X_{nk}\right)^2=\frac{1}{n}\sum_{k=1}^{c'_n-1}\left(\tilde{X}_{n,k}-X_{n,k}\right)^2 \le \frac{1}{n}\sum_{k=1}^{n}\left(\tilde{X}_{n,k}-X_{n,k}\right)^2 \xrightarrow{p} 0$$

- $Y_{n,k} = X_{n,k}$ for $c_n + 1 \leq k \leq n - U_n$, the set of observations of X that are larger than $\lim_{n\to\infty} \tilde{X}_{n,L_n}$ with probability one, so that

$$\frac{1}{n}\sum_{k=c_n+1}^{n-U_n} \left(Y_{nk} - X_{nk}\right)^2 = \frac{1}{n}\sum_{k=c_n+1}^{n-U_n} \left(X_{nk} - X_{nk}\right)^2 = 0$$

- $Y_{n,k} \in \Xi_n = \{\tilde{X}_{n,c'_n}, \ldots, \tilde{X}_{n,L_n}, X_{n,L_n+1}, \ldots, X_{n,c_n}\}$ such that

$$\begin{aligned} & n^{1/2}\left(Y_{n,k} - X_{n,k}\right) \xrightarrow{wp1} 0 \\ \Rightarrow \quad & n\left(Y_{n,k} - X_{n,k}\right)^2 \xrightarrow{wp1} 0 \end{aligned} \tag{1.56}$$

Let $c''_n = L_n + \frac{1}{2}n^{1/2}$. We have

$$\frac{c''_n}{n} = \frac{L_n + \frac{1}{2}n^{1/2}}{n} = \frac{L_n}{n} + \frac{1/2}{n^{1/2}} = \Phi(-3) + \frac{1/2}{n^{1/2}} + o(n^{-1/2})\,.$$

Because of a corollary of Serfling (1980, page 94), $\lim_{n\to\infty} X_{n,c''_n}$ will be larger than $\lim_{n\to\infty} \tilde{X}_{n,L_n}$ with probability one and this implies that X_{n,c''_n} will not be in Ξ_n for n large enough.

By analogy the same will follow for $\tilde{X}_{n,L_n-\frac{1}{2}n^{1/2}}$ so that the set Ξ_n will have at most $n^{1/2}$ elements for n sufficiently large. It then follows with equation (1.56) that

$$\frac{1}{n}\sum_{Y_{nk}\in\Xi_n} \left(Y_{nk} - X_{nk}\right)^2 \xrightarrow{wp1} 0$$

By analogy, the same can be shown for arranging Part 2 and 3 in ascending order. Hence we have the result

$$\frac{1}{n}\sum_{k=1}^{n}\left(Y_{nk} - X_{nk}\right)^2 \xrightarrow{p} 0\,. \quad \blacksquare$$

1.5.3.6 Proof of Theorem 2

Proof

We start by showing that

$$\left|\frac{\sum_{k=1}^{n} a_{nk}X_{nk}}{\sqrt{\sum_{k=1}^{n}\left(X_{nk}-\bar{X}\right)^2}} - \frac{\sum_{k=1}^{n} a_{nk}Y_{nk}}{\sqrt{\sum_{k=1}^{n}\left(Y_{nk}-\bar{Y}\right)^2}}\right| = \left|W_n^{1/2} - \tilde{W}_n^{1/2}\right| \xrightarrow{p} 0\,.$$

For the difference of the numerators, we can write

$$\frac{1}{n^{1/2}}\left|\sum_{k=1}^{n} a_{nk}X_{nk}-\sum_{k=1}^{n} a_{nk}Y_{nk}\right| \leq n^{-1/2}\sum_{k=1}^{n}|a_{nk}|\,|X_{nk}-Y_{nk}|$$

$$\leq \frac{1}{n^{1/2}}\left(\sum_{k=1}^{n} a_{nk}^2\right)^{1/2}\left(\sum_{k=1}^{n}(X_{nk}-Y_{nk})^2\right)^{1/2}. \qquad (1.57)$$

Because the condition $\sum_{k=1}^{n} a_{nk}^2 = 1$ is fulfilled and with application of Theorem 5, we have

$$\frac{1}{n^{1/2}}\left(\sum_{k=1}^{n} a_{nk}^2\right)^{1/2}\left(\sum_{k=1}^{n}(X_{nk}-Y_{nk})^2\right)^{1/2} = \left(\frac{1}{n}\sum_{k=1}^{n}(X_{nk}-Y_{nk})^2\right)^{1/2} \xrightarrow{p} 0\,. \qquad (1.58)$$

Hence with equation (1.57), we have

$$n^{-1/2}\left|\sum_{k=1}^{n} a_{nk}X_{nk}-\sum_{k=1}^{n} a_{nk}Y_{nk}\right| \xrightarrow{p} 0\,. \qquad (1.59)$$

For convenience, let

$$s_X^2 = \frac{1}{n}\sum_{k=1}^{n}\left(X_{nk}-\bar{X}\right)^2$$

$$s_Y^2 = \frac{1}{n}\sum_{k=1}^{n}\left(Y_{nk}-\bar{Y}\right)^2.$$

It is easy to see with some computation that

$$\left|\frac{\sum_{k=1}^{n} a_{nk}X_{nk}}{\sqrt{\sum_{k=1}^{n}\left(X_{nk}-\bar{X}\right)^2}}-\frac{\sum_{k=1}^{n} a_{nk}Y_{nk}}{\sqrt{\sum_{k=1}^{n}\left(Y_{nk}-\bar{Y}\right)^2}}\right| = n^{-1/2}\left|\frac{\sum_{k=1}^{n} a_{nk}X_{nk}}{\sqrt{s_X^2}}-\frac{\sum_{k=1}^{n} a_{nk}Y_{nk}}{\sqrt{s_Y^2}}\right|$$

$$=\left|\frac{n^{-1/2}\left(\sum_{k=1}^{n} a_{nk}X_{nk}-\sum_{k=1}^{n} a_{nk}Y_{nk}\right)}{\sqrt{s_X^2}}+\left(\frac{\sqrt{s_Y^2}-\sqrt{s_X^2}}{\sqrt{s_X^2}}\right)\left(n^{-1/2}\frac{\sum_{k=1}^{n} a_{nk}Y_{nk}}{\sqrt{s_Y^2}}\right)\right|$$

$$=\left|\frac{n^{-1/2}\left(\sum_{k=1}^{n} a_{nk}X_{nk}-\sum_{k=1}^{n} a_{nk}Y_{nk}\right)}{\sqrt{s_X^2}}+\left(\frac{\sqrt{s_Y^2}-\sqrt{s_X^2}}{\sqrt{s_X^2}}\right)\tilde{W}_n^{1/2}\right|$$

$$\leq\frac{n^{-1/2}\left|\sum_{k=1}^{n} a_{nk}X_{nk}-\sum_{k=1}^{n} a_{nk}Y_{nk}\right|}{\sqrt{s_X^2}}+\left|\frac{\sqrt{s_Y^2}-\sqrt{s_X^2}}{\sqrt{s_X^2}}\right|\tilde{W}_n^{1/2}\,.$$

It follows with equation (1.59) and Theorem 4 that

$$\frac{n^{-1/2}\left|\sum_{k=1}^{n} a_{nk}X_{nk}-\sum_{k=1}^{n} a_{nk}Y_{nk}\right|}{\sqrt{s_X^2}} \xrightarrow{p} 0 \quad \text{and}$$

$$\frac{\sqrt{s_Y^2}-\sqrt{s_X^2}}{\sqrt{s_X^2}} \xrightarrow{p} 0\,.$$

Hence because $0 \leq \tilde{W}_n^{1/2} \leq 1$ holds, we obtain

$$\left| W_n^{1/2} - \tilde{W}_n^{1/2} \right| \xrightarrow{p} 0 \,.$$

Consequently

$$\begin{aligned} \left| W_n - \tilde{W}_n \right| &= \left| 2W_n^{1/2} \left(W_n^{1/2} - \tilde{W}_n^{1/2} \right) - \left(W_n^{1/2} - \tilde{W}_n^{1/2} \right)^2 \right| \\ &\leq 2W_n^{1/2} \left| W_n^{1/2} - \tilde{W}_n^{1/2} \right| + \left| W_n^{1/2} - \tilde{W}_n^{1/2} \right|^2 \\ \Rightarrow \left| W_n - \tilde{W}_n \right| &\xrightarrow{p} 0 \,. \quad \blacksquare \end{aligned}$$

1.5.4 Simulations

To support the above demonstrations, we conduct some simulations.
We generated 10000 sequences from a standard normal distribution for different sample sizes ranging from 10 to 50000 observations to assess the convergence of the different distributions to the normal. We used 161 equidistant points between -8 and 8 and for each point x in this grid, we computed for every sample the values of $G_n(x)$, $\hat{G}_n(x)$, $F_n(x)$ and $\tilde{F}_n(x)$. This allows us to estimate the distribution of $\sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$, $\sup_{x\in\mathbb{R}} |G_n(x) - F(x)|$, $\sup_{x\in\mathbb{R}} \left|\hat{G}_n(x) - F(x)\right|$ and $\sup_{x\in\mathbb{R}} \left|\tilde{F}_n(x) - F(x)\right|$ from the maximum absolute errors over the grid. The asymptotic convergence of the first of these supremum distances to 0 is guaranteed by the Glivenko-Cantelli theorem. Theorem 1 derived in the previous subsections proves the same convergence for the last of the supremum distances, while the convergence of the other two terms has been obtained as auxiliary results, see Lemma 1 and 3. In particular, we use the average maximum errors to estimate the corresponding means. Furthermore we computed the average maximum squared errors over the grid. Figure 1.2 and 1.3 summarize the results of these simulations.

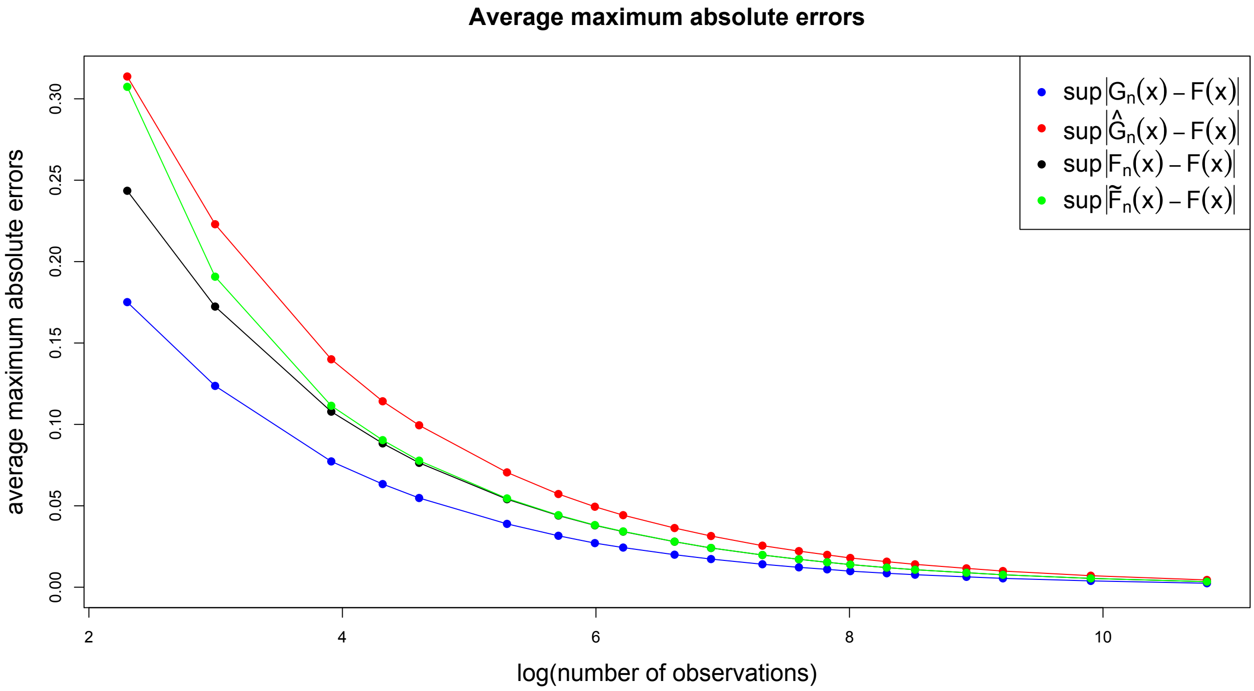


Figure 1.2: Average maximum absolute errors as a function of the sample size.

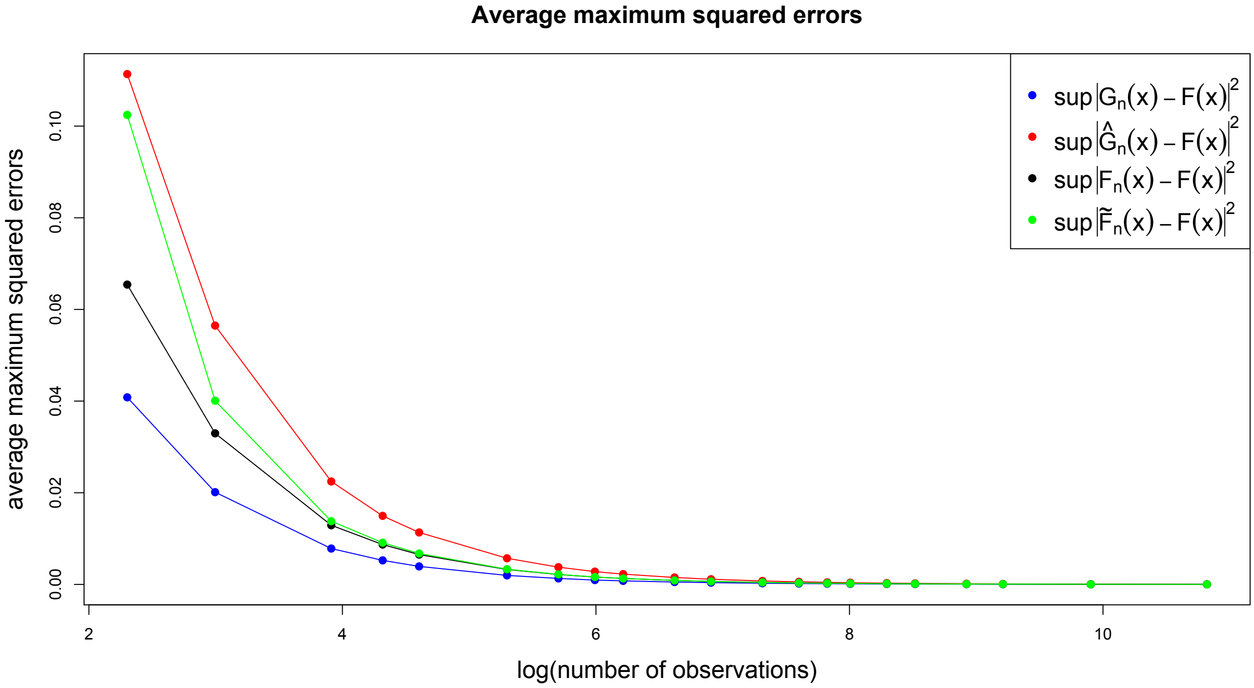


Figure 1.3: Average maximum squared errors as a function of the sample size.

Although $\hat{G}_n(x)$ converges slower than the other functions to F(x), this does not seem to affect the rate of convergence very much. To estimate the rate of convergence, we regress the logarithm of the average maximum absolute errors on the logarithm of the sample size. Table 1.1 presents the slopes and intercepts of the different regressions. All slopes are about -0.5, leading to the approximation

$$log(e_n) \approx -\frac{1}{2}log(n) + C \Leftrightarrow e_n \approx e^C n^{-1/2},$$

where e_n is either $\sup_{x\in\mathbb{R}} |G_n(x) - F(x)|$, $\sup_{x\in\mathbb{R}} \left|\hat{G}_n(x) - F(x)\right|$, $\sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$ or $\sup_{x\in\mathbb{R}} \left|\tilde{F}_n(x) - F(x)\right|$, n denotes the number of observations and e^C is a constant not depending on n. This means that the convergence rate seems to be the same for all functions because the supremum of the absolute difference between each of these distribution functions and $F(x)$ decreases at the ordinary rate $n^{-1/2}$ as the sample size n increases. The difference seems to be only the constant factor e^C. This can also be seen in Figure 1.4.

	Estimate	Std. Error	t value	p value
Regression of $log\left(\sup_{x\in\mathbb{R}}\left\|G_n(x)-F(x)\right\|\right)$ on $log(n)$				
Intercept	-0.5944	0.0041	-146.6	<2e-16
log(n)	-0.5013	0.0006	-875.8	<2e-16
Regression of $log\left(\sup_{x\in\mathbb{R}}\left\|\hat{G}_n(x)-F(x)\right\|\right)$ on $log(n)$				
Intercept	-0.0104	0.0031	-3.367	0.0032
log(n)	-0.4995	0.0004	-11147.861	<2e-16
Regression of $log\left(\sup_{x\in\mathbb{R}}\left\|F_n(x)-F(x)\right\|\right)$ on $log(n)$				
Intercept	-0.2647	0.0030	-88.32	<2e-16
log(n)	-0.5005	0.0004	-1182.75	<2e-16
Regression of $log\left(\sup_{x\in\mathbb{R}}\left\|\tilde{F}_n(x)-F(x)\right\|\right)$ on $log(n)$				
Intercept	-0.1389	0.030	-4.6	0.0002
log(n)	-0.5162	0.0043	-121.1	<2e-16

Table 1.1: Results of the regressions of the logarithm of $\sup_{x\in\mathbb{R}}|G_n(x)-F(x)|$, $\sup_{x\in\mathbb{R}}\left|\hat{G}_n(x)-F(x)\right|$, $\sup_{x\in\mathbb{R}}|F_n(x)-F(x)|$ and $\sup_{x\in\mathbb{R}}\left|\tilde{F}_n(x)-F(x)\right|$ on the logarithm of the number of observations.

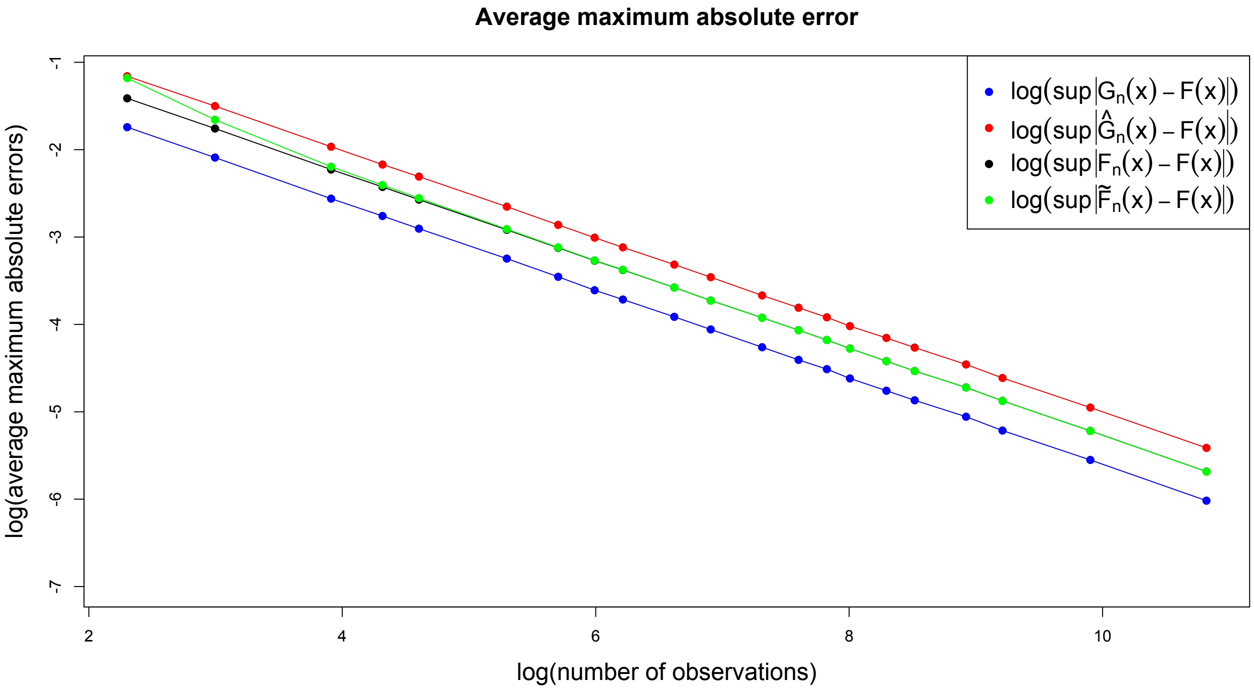


Figure 1.4: Logarithm of the average maximum absolute errors against the logarithm of the number of observations.

1.6 Comparison of the tests for normality

In this Section, we compare the robust tests for normality that we introduced in the previous sections.

1.6.1 Sizes of the tests

First we check the sizes of the different tests in finite samples. For this purpose, we simulate 10000 samples of size 100 from the standard normal distribution and compare the performance of the tests when the samples are clean and when they are contaminated with one positive and one negative outlier of the same magnitude equal to 7 standard deviations. We also consider contamination with 5 outliers (3 positive and 2 negative ones) of the same magnitude 7 standard deviations. The robust tests are compared to the results of the ordinary Shapiro-Wilk (SW) test. To avoid excessively long simulation times for the $TRIM_\alpha$ test, we only compute 1000 samples for it in the same manner as for the others. To be specific, all the tests implemented here except $TRIM_\alpha$ have similar computation times, while the computation time of the $TRIM_\alpha$ test is 9 to 1614 times higher, depending on the sample size, which ranges from 10 to 5000 observations, and on the procedure. The average computation times for all procedures are given in Table 1.2, from which one can see that in the case of samples of size 100, the computation time of $TRIM_\alpha$ is 55 to 537 time higher than that of the other test procedures.

	sample sizes								
	10	20	50	100	200	500	1000	2000	5000
SW	3.0	3.1	3.3	3.4	3.4	3.7	2.2	4.2	3.9
RSW	3.5	3.9	4.7	4.5	4.9	6.6	3.9	4.0	5.2
RSW_{AO}	5.5	5.6	5.4	6.0	7.5	11.0	21.4	32.5	65.1
RSW_{AB}	8.2	7.7	8.7	10.1	9.9	13.5	17.3	26.9	68.8
RSW_{AS}	4.8	5.8	5.0	6.4	6.7	6.0	4.9	4.8	10.9
SJ	2.8	2.1	1.8	1.5	2.4	2.1	3.8	2.3	1.9
JB	1.0	0.4	0.7	1.1	0.6	0.4	1.1	1.6	1.0
$MC1$	4.1	3.1	4.2	2.0	3.0	2.4	5.6	6.1	7.2
$MC2$	5.3	6.3	6.0	4.9	5.3	5.9	5.5	7.1	8.7
$MC3$	5.6	6.7	6.1	6.8	7.9	7.3	4.7	7.3	15.3
RJB	2.2	2.0	1.7	2.0	0.7	1.2	1.1	1.5	1.7
$TRIM_{\alpha}$	573.4	540.7	544.1	561.1	557.7	586.0	584.0	599.0	613.6

Table 1.2: Average computation times (in milliseconds) for each procedure for sample sizes between 10 and 5000.

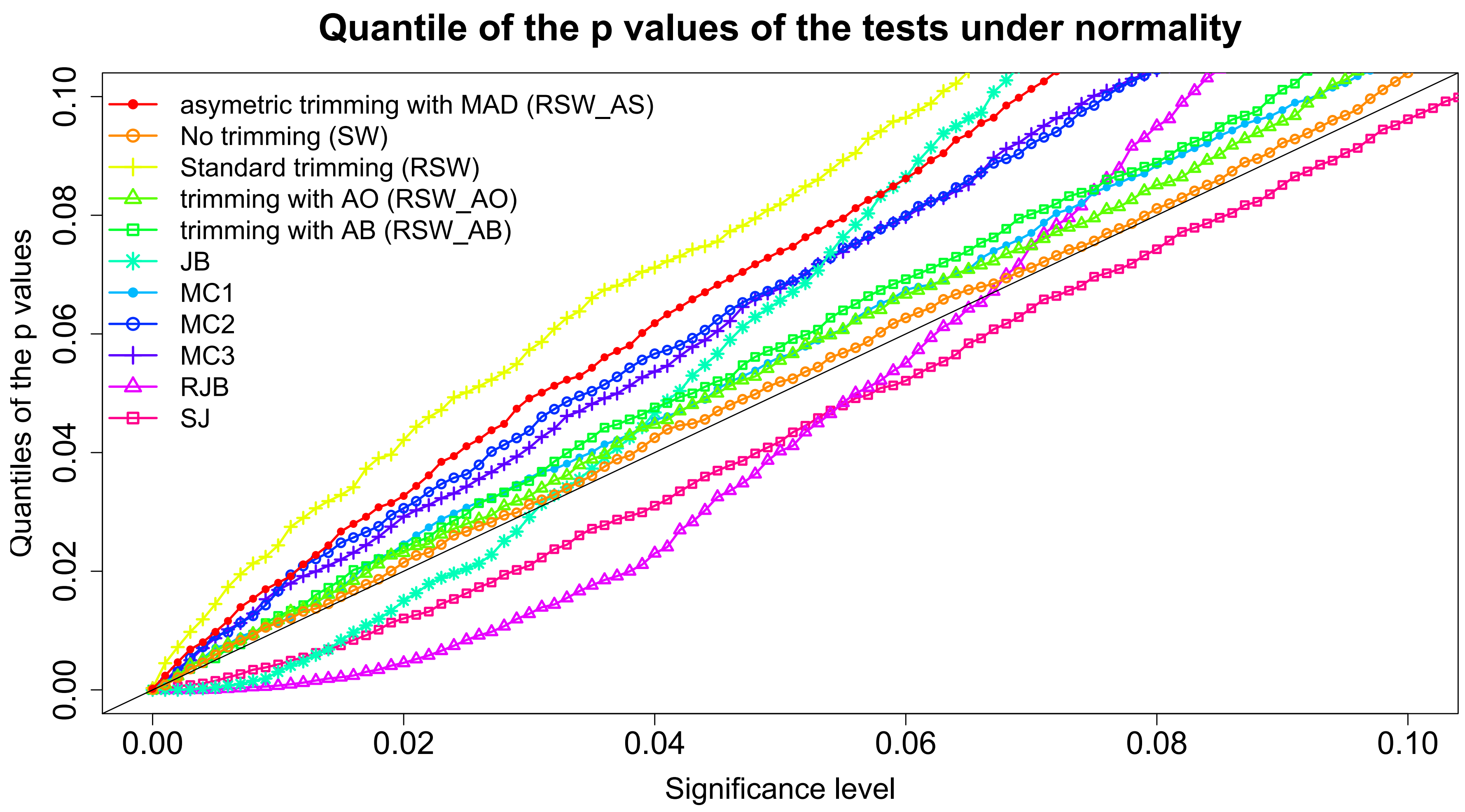


Figure 1.5: QQ Plot of p values of robust tests for normality for 1000 clean standard normal samples of length 100.

Figure 1.5 presents the quantiles of the p values of the different tests as a function of the significance level in case of clean normal samples. The $TRIM_{\alpha}$ test does not appear in the graphic because it needs a threshold in order to compute p values, and because the hypothesis that are tested in its case are not the same as for the other tests. More precisely, for all the other procedures, the null hypothesis tested is that the sample is normally distributed, as opposed to the $TRIM_{\alpha}$ test, which assumes non normality under the null hypothesis, see Section 1.3.4. Therefore, it is difficult to compare the computed p values. From this type of graphic, we can see whether the test is conservative or liberal. A test is said to be conservative, when the probability of a type 1 error is less than the nominal significance level considered. The test is liberal if the opposite is true. So if p and α are respectively the p value of the observed test statistic and the nominal significance level, then a test is said to be conservative at level α, if

$$\mathit{Probability}(p \leq \alpha) < \alpha.$$

Hence, if q_{α} is the α quantile of the p value of the test and $q_{\alpha} > \alpha$, then the test is conservative.
Figure 1.5 reveals that the majority of the tests is conservative, while the SJ test is liberal. Note that the JB and the RJB test are liberal for small values of α and become conservative afterwards. At a significance level of 5%, the percentages of rejection of the null hypothesis of normality are given in Table 1.3. Note that we have used simulations for clean data to deduce a threshold for the $TRIM_{\alpha}$ test. That is why the percentage of rejection of normality is exactly 5%. We will use the same threshold $\Delta_0^2 = 0.014647$ for the rest of the comparisons. Similarly, Figure 1.6 and 1.7 illustrate the p values of the tests in the presence of 2 and 5 outliers, respectively. We note a complete breakdown of the SW, the JB, the RJB and the SJ test, since they reject normality in the presence of a few outliers. This is not surprising because the SW and the JB test are not robust and the RJB and the SJ tests are not robust against outliers, but are designed to detect heavy tails. We also notice a breakdown, although not complete, of the RSW_{AO} for significance levels less than about 0.27 in the presence of 2 outliers and 0.36 in the case of 5 outliers. For this reason it does not appear in the graphics. Because the most commonly

Test	Rejection of normality (in %)	Test	Rejection of normality (in %)
SW	4.88	JB	4.24
RSW	2.44	RJB	5.68
RSW_{AO}	4.43	$MC1$	4.53
RSW_{AB}	4.04	$MC2$	3.82
RSW_{AS}	3.02	$MC3$	3.87
SJ	5.91	$TRIM_\alpha$	5.00

Table 1.3: Percentage of rejection of the null hypothesis of normality at a significance level of 5% for samples of length 100.

used significance level is typically chosen as 5% and very rarely higher than 10%, we can consider the RSW_{AO} to break down. The breakdown of the RSW_{AO} is due to the fact that the adjusted outliers method does not detect outliers efficiently enough for the trimming to replace them correctly. This leads to a higher rejection rate of the normality hypothesis. The other tests reveal no significant changes and demonstrate good robustness to the two outliers. The percentage of rejection of the null hypothesis in presence of two outliers are summarized in Table 1.4: we note the breakdowns that we have already noticed in Figures 1.6 and 1.7. But we also notice that the presence of the outliers affects the ability of the $TRIM_\alpha$ test to detect normality.

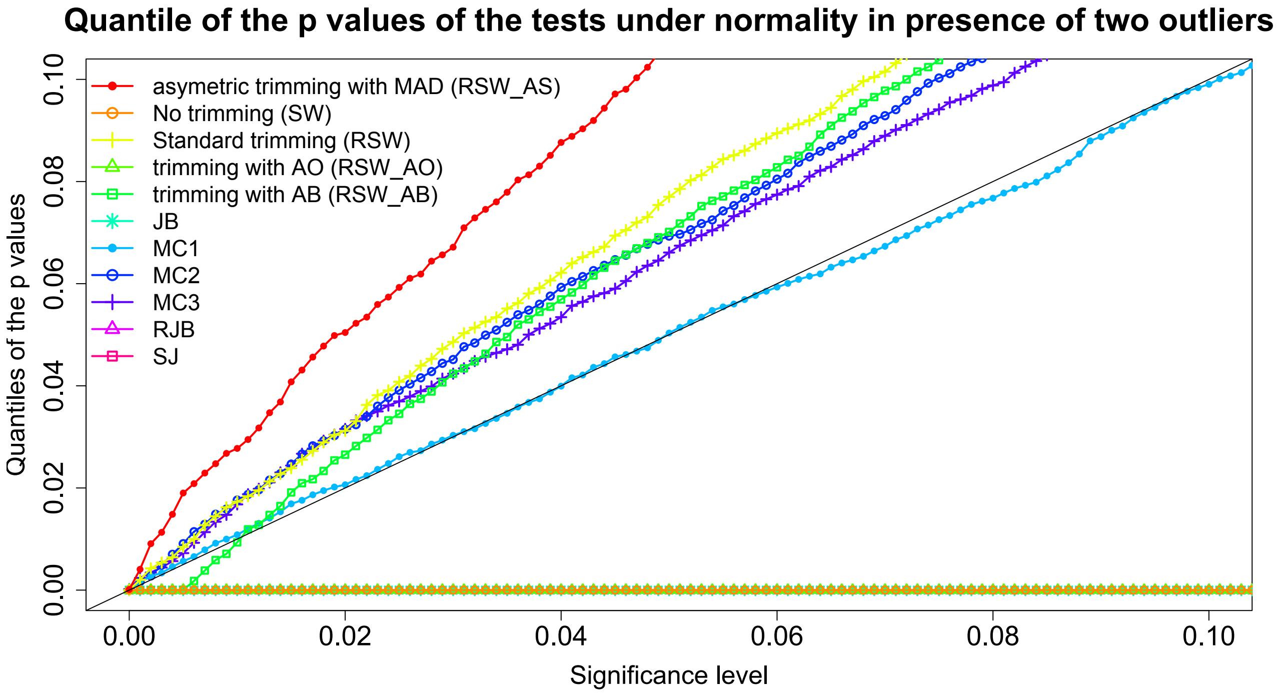


Figure 1.6: QQ Plot of p values of robust tests for normality for contaminated standard normal samples of length 100 with two outliers of magnitude 7 standard deviations each.

Test	Rejection of normality (in %)		Test	Rejection of normality (in %)	
	2 outliers	5 outliers		2 outliers	5 outliers
SW	100	100	JB	100	100
RSW	1.77	2.13	RJB	100	100
RSW_{AO}	28.27	37.52	$MC1$	4.7	4.6
RSW_{AB}	3.9	3.97	$MC2$	3.57	3.52
RSW_{AS}	2.07	2.03	$MC3$	3.84	3.86
SJ	100	100	$TRIM_{\alpha}$	11.40	35.25

Table 1.4: Percentage of rejections of the null hypothesis of normality in presence of outliers of magnitude 7 standard deviations at a significance level of 5% for samples of length 100.

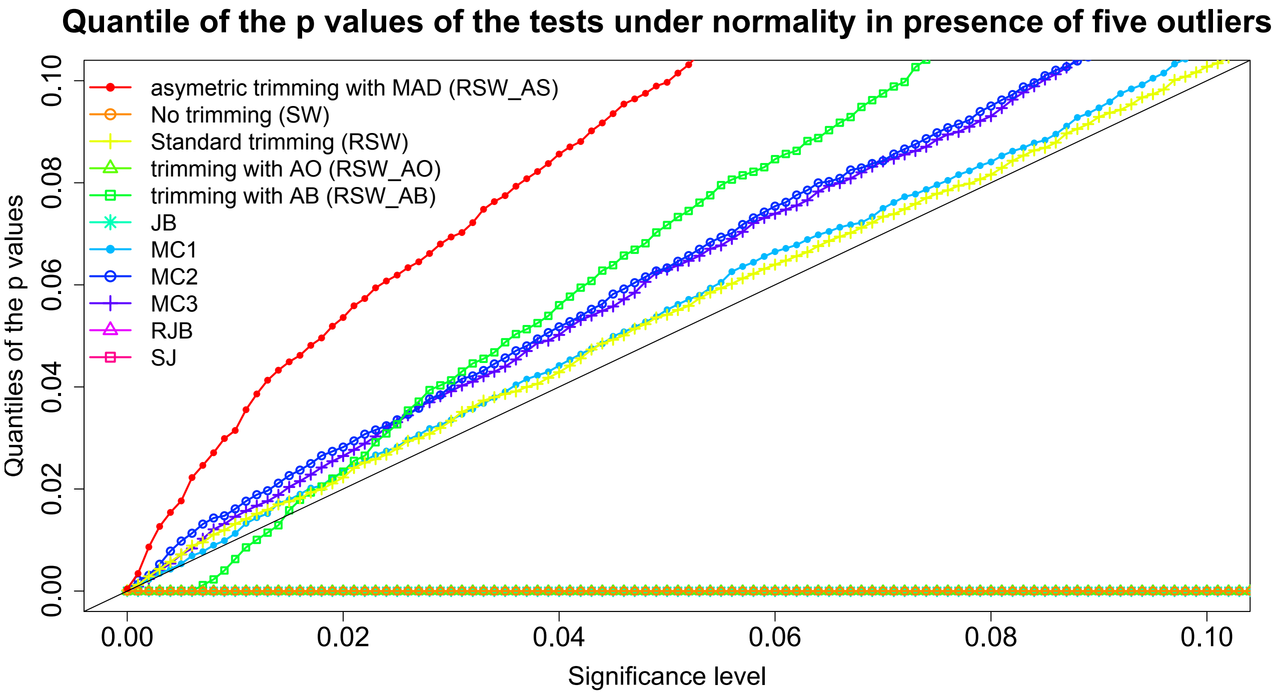


Figure 1.7: QQ Plot of p values of robust tests for normality for contaminated standard normal samples of length 100 with five outliers of magnitude 7 standard deviations each.

1.6.2 Power of the robust tests

In this subsection, we compare the power of the robust tests for normality with the SW test. To investigate different departures from normality, we simulate 10000 samples of length 100 from each of the following distributions:

- The χ^2 distribution with 2 or 10 degrees of freedom
- The t distribution with 2, 3, 5 or 10 degrees of freedom
- The inverse Box-Cox with parameter λ applied to a normal distribution with mean 7 and variance 1, where $\lambda \in \{0, 0.25, 0.5, 0.75\}$
- The inverse Box-Cox with parameter λ applied to a normal distribution with mean 7 and variance 1 with two outliers with the values 1 and 13 (i.e. 6 standard deviations) before inverse Box-Cox transformation, where $\lambda \in \{0, 0.25, 0.5, 0.75\}$
- The inverse Box-Cox with parameter λ applied to a normal distribution with mean 7 and variance 1 in the presence of five outliers (3 positive and 2 negative outliers) with the values 1 and 13 before inverse Box-Cox transformation, where $\lambda \in \{0, 0.25, 0.5, 0.75\}$

The results are summarized in Tables 1.5 to 1.8.

From Table 1.5, we can deduce that all tests perform similarly well except the robust tests based on the medcouple, which have poor power against these alternatives. The RSW_{AS} is more powerful than its symmetric version RSW in case of a χ^2 distribution, but has little power for detecting a t-distribution. In case of a t-distribution the JB, the RJB and the SJ test show the best results and they outperform the SW and the $TRIM_\alpha$ test. This is not surprising because their power is directed towards such departures from normality.

Tables 1.6, 1.7 and 1.8 summarize the results of the power study in case of departures corresponding to an inverse Box-Cox transformation of a normal distribution. For $\lambda = 1$, we have a normal distribution, and as the difference between λ and 1 grows the distribution

Test	χ^2_2	χ^2_{10}	t_2	t_3	t_5	t_{10}
RSW_{AS}	100.0	79.8	10.8	7.5	5.5	3.9
SW	100.0	90.4	98.6	87.4	56.8	23.5
RSW	100.0	67.7	8.4	5.6	4.0	2.6
RSW_{AO}	100.0	85.8	83.3	61.8	34.1	14.0
RSW_{AB}	100.0	75.3	26.1	18.8	12.3	7.0
JB	100.0	81.3	98.8	90.4	64.3	31.0
$MC1$	85.2	24.0	7.6	6.3	5.5	5.5
$MC2$	47.6	8.9	19.7	9.3	6.0	5.1
$MC3$	95.7	24.1	18.7	9.1	6.0	5.3
RJB	100.0	74.2	99.3	92.5	67.7	32.9
SJ	87.3	22.2	99.1	90.0	57.7	21.1
$TRIM_\alpha$	100.0	80.5	89.5	63.1	23.0	11.2

Table 1.5: Power (in %) of the tests for normality at a significance level of 5%.

departs exponentially fast from normality. Therefore we expect the power of the tests to decrease as the value of λ tends to 1. Note that almost all performances are similar except the one of the $MC2$ test. However, in the presence of outliers, the SW, the JB, the RJB and the SJ test break down.

As an illustration of the behaviour of the tests with a skewed distribution, we represent the power of the tests for normality applied to a log-normal distribution ($\lambda = 0$) as a function of the significance level. The results of 100000 samples are seen in Figures 1.8 to 1.10. In the case of the $TRIM_\alpha$ test, we used 1000 samples. The robust tests based on the medcouple have the smallest power, with the $MC2$ test being the least powerful.

Test	$\lambda = 0$	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 0.75$
RSW_{AS}	100.0	65.0	13.0	4.5
SW	100.0	81.8	21.0	7.3
RSW	100.0	51.2	8.9	3.5
RSW_{AO}	100.0	74.9	18.3	6.6
RSW_{AB}	100.0	61.8	13.1	5.6
JB	100.0	72.9	18.8	6.4
$MC1$	94.6	20.1	6.9	5.0
$MC2$	48.1	8.2	5.4	4.8
$MC3$	99.0	20.1	6.7	4.9
RJB	100.0	66.9	17.5	6.5
SJ	99.6	20.0	7.0	5.3
$TRIM_{\alpha}$	100.0	70.0	18.2	7.9

Table 1.6: Power (in %) of the tests for normality for the normal distribution with mean 7 and variance 1 transformed with the inverse Box-Cox with parameter λ.

Test	$\lambda = 0$	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 0.75$
RSW_{AS}	100.0	40.1	6.5	2.9
SW	100.0	100.0	100.0	100.0
RSW	100.0	20.0	2.9	2.2
RSW_{AO}	100.0	58.8	32.5	41.5
RSW_{AB}	100.0	46.8	10.1	6.5
JB	100.0	100.0	100.0	100.0
$MC1$	95.0	20.4	7.1	5.9
$MC2$	50.2	7.6	5.5	5.1
$MC3$	99.1	20.3	7.1	5.6
RJB	100.0	100.0	100.0	100.0
SJ	100.0	100.0	100.0	100.0
$TRIM_{\alpha}$	99.3	64.9	26.8	14.2

Table 1.7: Power (in %) of the tests for normality for the normal distribution with mean 7 and variance 1 transformed with the inverse Box-Cox with parameter λ in presence of two outliers.

Test	$\lambda = 0$	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 0.75$
RSW_{AS}	100.0	33.3	7.5	2.7
SW	100.0	100.0	100.0	100.0
RSW	100.0	14.4	3.8	3.2
RSW_{AO}	100.0	56.0	58.1	55.2
RSW_{AB}	100.0	40.6	13.2	6.8
JB	100.0	100.0	100.0	100.0
$MC1$	96.1	22.4	7.8	5.6
$MC2$	51.6	6.5	4.3	3.8
$MC3$	99.6	20.8	6.9	4.3
RJB	100.0	100.0	100.0	100.0
SJ	100.0	100.0	100.0	100.0
$TRIM_{\alpha}$	100.0	67.6	39.6	33.6

Table 1.8: Power (in %) of the tests for normality for the normal distribution with mean 7 and variance 1 transformed with the inverse Box-Cox with parameter λ in presence of three positive outliers and two negative outliers each of magnitude 6 standard deviations.

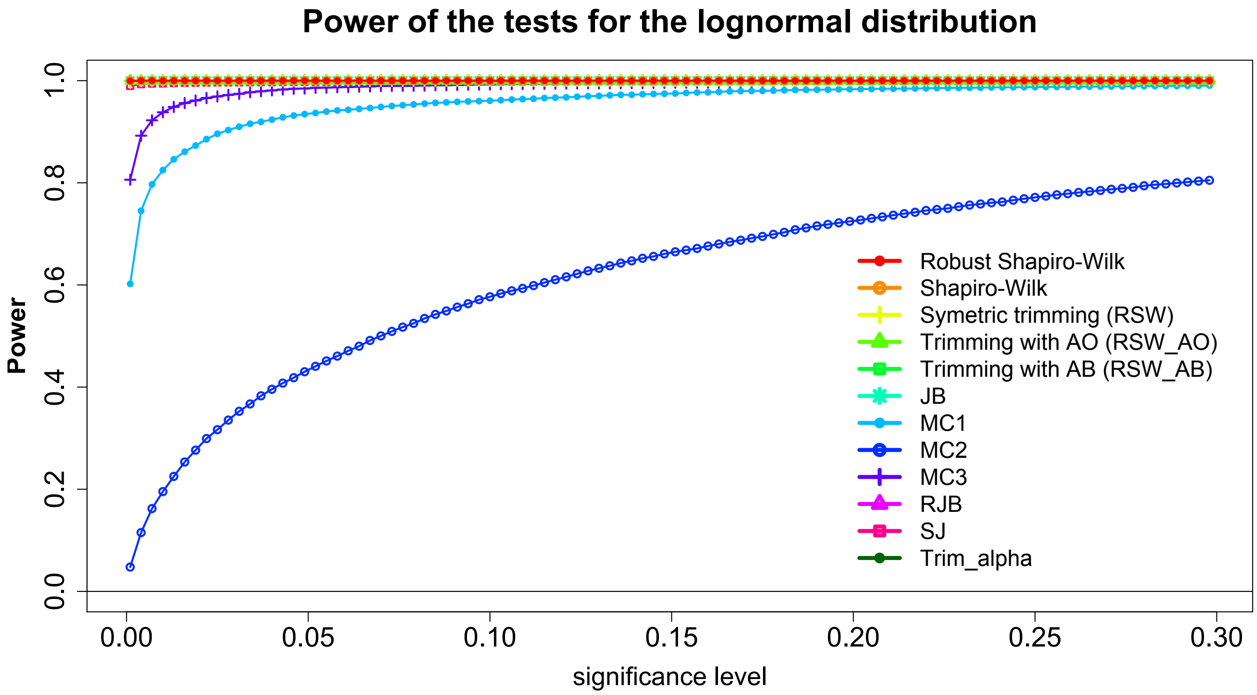


Figure 1.8: Power of the normality tests when the samples follow a log-normal distribution.

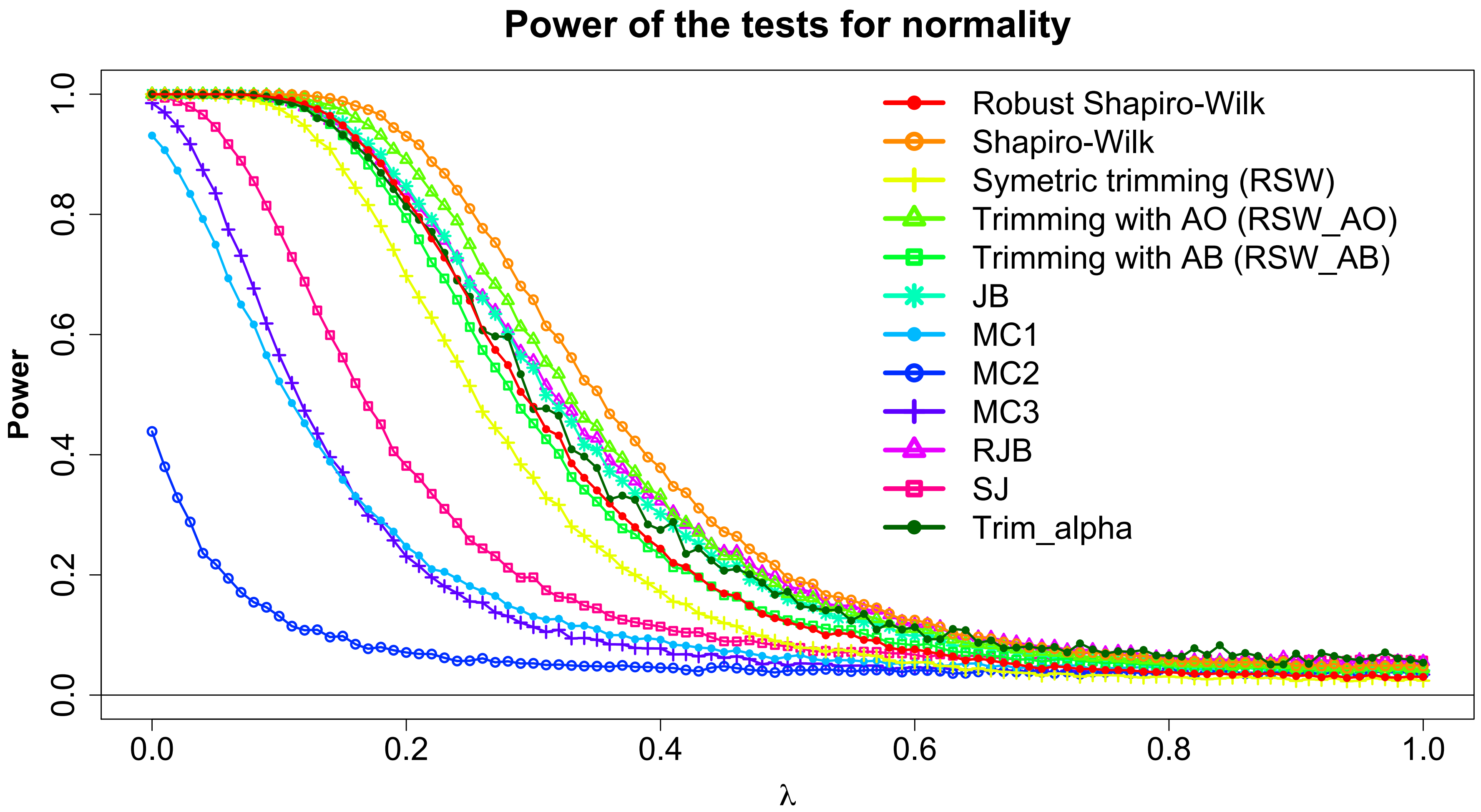


Figure 1.9: Power of the normality tests as a function of the transformation parameter λ of an inverse Box-Cox transformed normal distribution if there are no outliers.

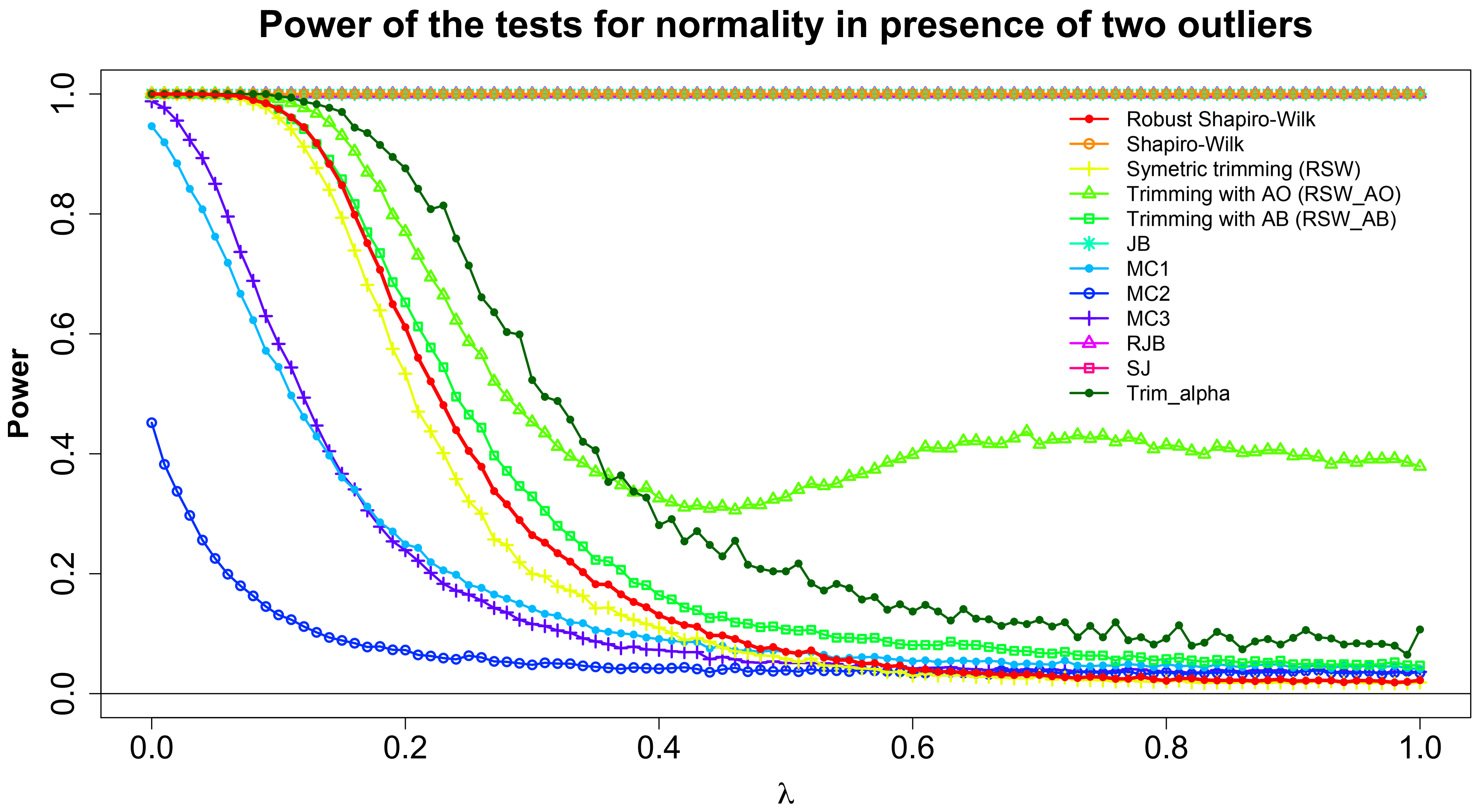


Figure 1.10: Power of the normality tests as a function of the transformation parameter λ of an inverse Box-Cox transformation of a normal distribution in presence of one negative and one positive outlier of magnitude 6 standard deviations each.

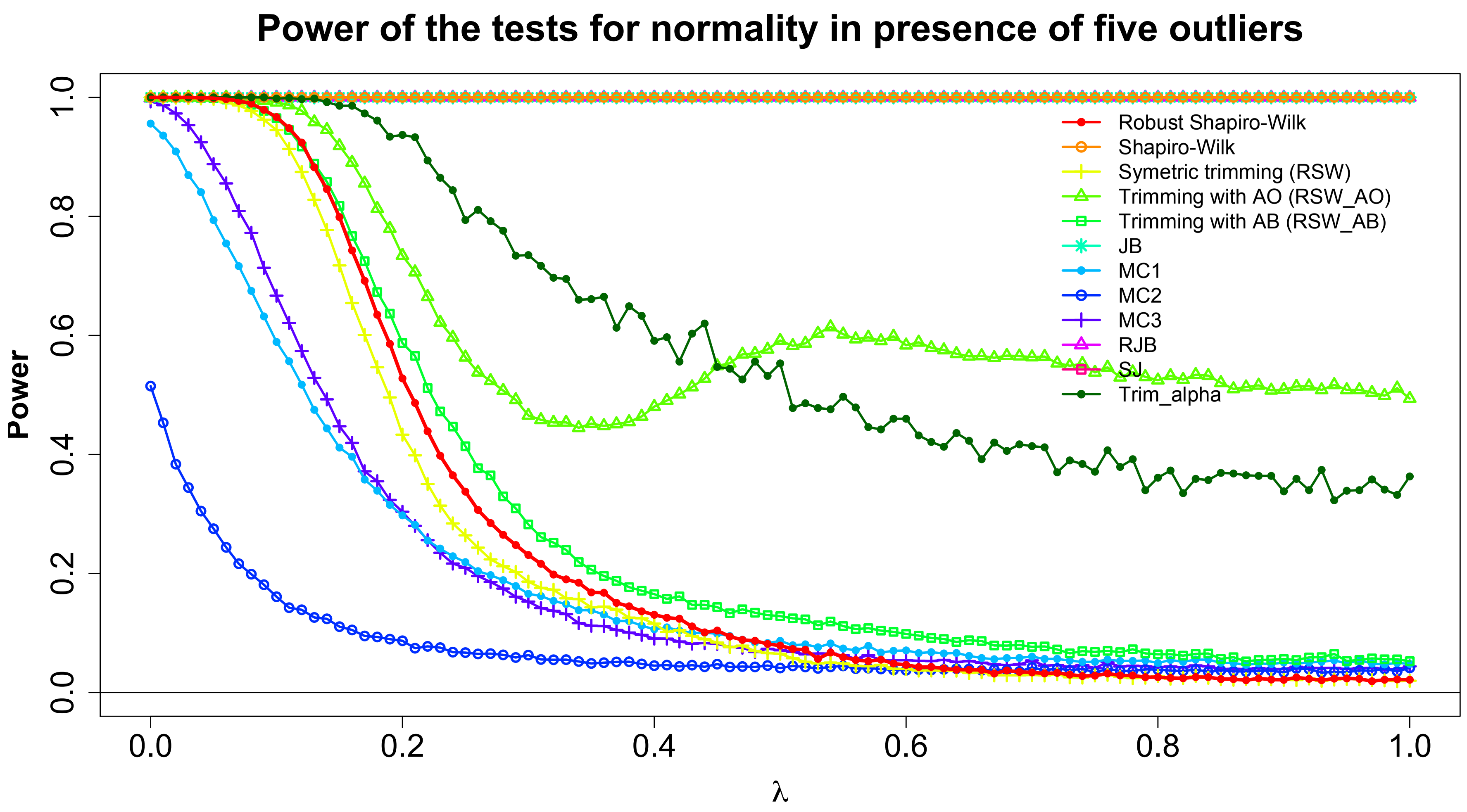


Figure 1.11: Power of the normality tests as a function of the transformation parameter λ of an inverse Box-Cox transformation of a normal distribution in presence of two negative and three positive outliers of magnitude 6 standard deviations each.

Figures 1.9, 1.10 and 1.11 show the power of the tests for normality as a function of the transformation parameter λ in the case of clean samples or samples contaminated with 2 and 5 outliers of size 6 standard deviations each, respectively. The RSW_{AS}, the RSW and the RSW_{AB} show the best performances as they are robust against the outliers and meet the expectations in terms of power. We also note that the RSW_{AS} seems to perform slightly better than the other two robust tests. The SW, the JB, the RJB and the SJ test are not robust and we also notice that in the presence of outliers, the power of the RSW_{AO} test tends to stabilize at a certain level, mainly because it fails to detect the outliers efficiently. The $TRIM_\alpha$ test shows good power, but when we approach normality (as the value of λ increases and gets closer to one), the rejection rate of the normality assumption for the $TRIM_\alpha$ test is very high compared to the other tests and it becomes larger as the number of outliers in the data increases, as we can already note in Table 1.8. Its curve is not smooth due to the fact that we used only 1000 samples per value of λ for it, instead of 10000 as for the other tests.

1.7 Conclusion

The aim of this chapter is to test whether the majority of a data sequence follows a normal distribution, in other words approximate normality. This feature is particularly interesting, when the data is contaminated with a few outliers. We opted to robustify the Shapiro-Wilk test for normality and hopefully derive a test that outperforms its competitors in this context, because the Shapiro-Wilk test is one of the most powerful test of normality. The robust test RSW_{AS} that we have constructed actually meets our expectations in terms of power and we show that under the null hypothesis of normality, the difference between the new robust test statistic and the Shapiro-Wilk test statistic converges in probability to 0. This in turn implies that the asymptotic distribution of the new robust test statistic is the same as that of the Shapiro-Wilk test statistic. Intensive simulations illustrate not only that our robust test (RSW_{AS}) is not time consuming, but also that it is more robust than the other robust tests of normality in presence of two and five outliers, when the true distribution of the data is Gaussian. Another appealing feature of the robust Shapiro-Wilk test is its behaviour in presence of outliers when the inverse Box-Cox transformation family is considered. In this case, it behaves better than the others robust tests in terms of power. This feature will be very useful for deriving a suitable robust Box-Cox transformation in the next chapter.

Chapter 2

Robust Box-Cox transformation

2.1 Introduction

Many statistical methods work under the assumption that an underlying process is normally distributed. But in practice, this is not always the case. Because of this, there are some methods to transform non-normal data into normally distributed ones. A popular method is the Box-Cox transformation $T_\lambda\colon y \to y^{(\lambda)}$, where $y^{(\lambda)}$ is given by the following formula:

$$y^{(\lambda)} = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log y, & \text{if } \lambda = 0\,. \end{cases} \tag{2.1}$$

We will restrict our analysis to this transformation family.

By its definition, the Box-Cox transformation can only be applied to positive data. Bickel & Doksum (1981) extended the definition to negative data as follows:

$$y^{(\lambda)} = \begin{cases} \dfrac{|y|^\lambda sgn(y) - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log |y|, & \text{if } \lambda = 0, \end{cases} \tag{2.2}$$

where *sgn* is the function returning the sign of its argument. This definition (2.2) will be used in the rest of our analysis. Even though the transformation was extended to account for negative values, there is no reason to think the transformed data will actually follow a normal distribution, only approximative normality is achieved.

There has been a variety of proposals on how to estimate the parameter λ of the Box-Cox family. Box & Cox (1964) proposed maximum-likelihood and Bayes estimates of λ and also developed a likelihood ratio test for null hypothesis of the form $H_0 : \lambda = \lambda_0$.

Andrews (1971) showed the well known fact that the maximum-likelihood estimate of λ is sensible to outliers and developed the so-called significance method based on the F-test, which he illustrated to be insensitive to one outlier in his example.

Atkinson (1973) emphasized the fact that Andrews (1971) omitted the Jacobian of the transformation in his analysis and found in a Monte Carlo simulation that the likelihood ratio test is uniformly more powerful than Andrew's F-test. Further, in a Monte Carlo simulation, Atkinson's proposal to test $H_0 : \lambda = \lambda_0$ is found to be equivalent to the likelihood ratio test under normality.

Gaudard & Karson (2000) introduced estimates of λ based on the optimization of the W statistic of the Shapiro-Wilk test and two measures of symmetry, the kurtosis and the skewness. In their simulations the estimate based on the optimization of the Shapiro-Wilk statistic ($\hat{\lambda}_{SW}$) outperformed the maximum-likelihood estimate and the estimates based on the third and fourth moments. They also note that the variance of $\hat{\lambda}_{SW}$ is slightly higher than that of the maximum likelihood estimate $\hat{\lambda}_{ML}$.

Another proposal is the method of percentile introduced in Hinkley (1975) and compared with the maximum-likelihood estimator by Chung et al. (2007). The method of percentile outperforms $\hat{\lambda}_{ML}$ and is recommended for its simplicity. For a more detailed review on the Box-Cox transformation see Saskia (1992).

Because data is often contaminated with outliers, our interest is directed towards a robust estimator of λ, which all the above mentioned procedures fail to deliver. In the following section, we will present the maximum-likelihood estimate in more details. Section 3 presents an alternative algorithm to compute the maximum-likelihood estimator and a new estimator is introduced. Section 4 will be devoted to robust estimators of the transformation parameter and we use our new robust test of normality as a robust alternative to estimate λ. In a final section, we will compare the robust estimators in simulations.

2.2 Maximum-Likelihood estimation

Let $Y_1, \dots, Y_n$ be continuous non negative i.i.d. random variables. These variables are assumed to be positive for the Box-Cox transformation to be well defined. Using the maximum-likelihood approach, we assume that the following relationship holds:

$$Y_i^{(\lambda)} = \mu + e_i, \quad i = 1, 2, \dots, n \tag{2.3}$$

where μ is a constant and e_i is a normally distributed error term with mean 0 and variance σ^2. The Jacobian of the transformation from y_i to $y_i^{(\lambda)}$ is $y_i^{\lambda-1}$, so that the log-likelihood of the observed sample is given by

$$L(\lambda) = -(n/2)\log 2\pi - (n/2)\log \sigma^2 - (2\sigma^2)^{-1}\sum_{i=1}^{n}[y_i^{(\lambda)} - \mu]^2 + (\lambda - 1)\sum_{i=1}^{n}\log y_i\,. \tag{2.4}$$

This log-likelihood function should be maximized with respect to σ^2, μ and λ.

$$\frac{\partial L(\lambda)}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}\left(y_i^{(\lambda)} - \mu\right)^2 \tag{2.5}$$

$$\frac{\partial L(\lambda)}{\partial \mu} = \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i^{(\lambda)} - \mu\right) \tag{2.6}$$

$$\frac{\partial L(\lambda)}{\partial \lambda} = \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\left(y_i^{(\lambda)} - \mu\right)\left(\left(y_i^{(\lambda)} + \frac{1}{\lambda}\right)\log y_i - \frac{1}{\lambda}y_i^{(\lambda)}\right)\right) + \sum_{i=1}^{n}\log y_i\,. \tag{2.7}$$

By setting the derivatives in (2.5) and (2.6) equal to zero we get:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i^{(\lambda)} - \mu\right)^2 \tag{2.8}$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}y_i^{(\lambda)}\,. \tag{2.9}$$

Setting the derivative (2.7) to zero would be cumbersome. Instead of this, we use the expressions of $\hat{\sigma}^2$ and $\hat{\mu}$ in (2.8) and (2.9) in the log-likelihood to obtain a profile likelihood that is much easier to be maximised with respect to the parameter λ only.

Chung et al. (2007) propose an alternative algorithm for the computation of the maximum-likelihood estimator of λ. Assuming model (2.3), we obtain the log-likelihood of equation (2.4).

If we divide the observed sample by its geometric mean scale $g = \exp(n^{-1}\sum \log y_i)$, the last term of the likelihood disappears. This implies that we can obtain the scaled version of the parameter λ for the Box-Cox transformation, and then retrieve λ for the original unscaled transformation. Correspondingly, if we denote $y_i^* = \frac{y_i}{g}$ then it follows:

$$\begin{aligned}
y_i^{*(\lambda)} &= \frac{(\frac{y_i}{g})^\lambda - 1}{\lambda} = \frac{y_i^\lambda - g^\lambda}{\lambda g^\lambda} = \frac{1}{g^\lambda}\left(\frac{y_i^\lambda - 1}{\lambda} - \frac{g^\lambda - 1}{\lambda}\right) = \frac{1}{g^\lambda}\left(y_i^{(\lambda)} - g^{(\lambda)}\right) \\
&= \frac{1}{g^\lambda}\left(\mu + e_i - g^{(\lambda)}\right) = \frac{\mu - g^{(\lambda)}}{g^\lambda} + \frac{e_i}{g^\lambda} \\
y_i^{*(\lambda)} &= \mu^* + e_i^*, \qquad (2.10)
\end{aligned}$$

where

$$\begin{aligned}
\mu^* &= \frac{\mu - g^{(\lambda)}}{g^\lambda} \quad \text{and} \\
e_i^* &= \frac{e_i}{g^\lambda}\,.
\end{aligned}$$

The likelihood of the scaled sample is now given by:

$$L^*(\lambda) = -(n/2)\log 2\pi - (n/2)\log \sigma^{*2} - (2\sigma^{*2})^{-1}\sum_{i=1}^{n}[y_i^{*(\lambda)} - \mu^*]^2,$$

where σ^{*2} is the variance of e_i^{*2}. The maximum-likelihood estimators of the parameters are derived by maximizing the log-likelihood $L^*(\lambda)$ with respect to μ, σ^{*2} and λ. Setting the corresponding derivatives to zero yields the following equations:

$$\sum_{i=1}^{n} e_i^* = 0 \qquad (2.11)$$

$$\sigma^{*2} = \frac{1}{n}\sum_{i=1}^{n} e_i^{*2} \qquad (2.12)$$

$$\left(\frac{1}{\lambda^2\sigma^{*2}}\right)\sum_{i=1}^{n} e_i^*[\lambda y_i^{*\lambda}\log y_i - (y_i^{*(\lambda)} - 1)] = 0\,. \qquad (2.13)$$

We can derive the maximum-likelihood estimator of μ^* from (2.11) as:

$$\widehat{\mu^*} = \frac{1}{n}\sum_{i=1}^{n} y_i^{*(\lambda)}\,. \qquad (2.14)$$

Then from (2.12), we can write:

$$\widehat{\sigma^*}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i^{*(\lambda)} - \widehat{\mu^*}\right)^2 . \tag{2.15}$$

After some transformations equation (2.13) becomes:

$$\sum_{i=1}^{n}\widehat{e_i^*}\left(y_i^{*\lambda}\log(y_i) - \widehat{e_i^*}\right) = 0, \tag{2.16}$$

where $\widehat{e_i^*} = y_i^{*(\lambda)} - \widehat{\mu^*}$. If we set $f_i(\lambda) = y_i^{*\lambda}\log(y_i)$, we can use a first order Taylor expansion to approximate $f_i(\lambda)$ around a starting value λ. This method will provide a unique solution that is as good as the one obtained with higher order Taylor expansions. Rewriting equation (2.16) and solving for λ yields the following equation

$$\hat{\lambda} = \frac{\sum_{i=1}^{n}\widehat{e_i^*}^2 + \sum_{i=1}^{n}\widehat{e_i^*}[\lambda f'(\lambda) - f(\lambda)]}{\sum_{i=1}^{n}\widehat{e_i^*}f'(\lambda)} . \tag{2.17}$$

The algorithm to compute the maximum-likelihood estimators can thus be conducted as follows:

1. Choose an initial value λ
2. Transform the original sample using the Box-Cox transformation with this parameter λ, then compute $\widehat{\mu^*}$ and $\widehat{\sigma^*}^2$ according to equation (2.14) and (2.15), respectively
3. Compute a new value λ_c of λ using equation (2.17)
4. Check whether the difference between λ_c and λ is less than a predetermined precision level. If not, replace λ by λ_c and iterate between step 1 and step 3 until the difference between λ_c and λ is smaller than the precision level before.

2.3 Robust estimators of λ

All the methods mentioned previously fail to produce estimates which are robust against outliers. Due to the fact that we are interested in a transformation to normality in the presence of outliers, we consider in this section some robust alternatives to estimate the transformation parameter, and only these procedures will be considered in future comparisons due to their robustness.

2.3.1 M-estimates of λ

Carroll (1980) proposed a method based on the idea of M-estimation of Huber (1964). The idea is to replace the normal theory likelihood given by Box and Cox by the density with normal centre-exponential tails

$$L(\mu, \sigma, \lambda) = \sigma^{-n} \prod_{i=1}^{n} \exp\left[-\rho\left(\frac{y_i^{(\lambda)} - \mu}{\sigma}\right) + (\lambda - 1)\ln y_i\right] \tag{2.18}$$

where ρ is the Huber function given by

$$\rho(x) = \begin{cases} \frac{1}{2}x^2, & \text{if } |x| \leq k \\ k\left(|x| - \frac{1}{2}k\right), & \text{otherwise} \end{cases}$$

and k is chosen by the statistician. Common values are $k = 1.5$ and $k = 2$.
For a fixed value of λ, the estimation procedure works in the following steps :

1. Take an initial value of σ and estimate μ by solving

$$\sum_{i=1}^{n} \psi\left(\frac{y_i^{(\lambda)} - \mu}{\sigma}\right) = 0\,, \tag{2.19}$$

   where ψ is the first derivative of the Huber-ρ function ρ.

2. Then estimate σ by solving

$$(n-1)^{-1} \sum_{i=1}^{n} \psi^2\left(\frac{y_i^{(\lambda)} - \mu}{\sigma}\right) = E\left(\psi^2(Z)\right)\,, \tag{2.20}$$

   where Z is a standard normal random variable.

For a given λ denote the estimates as $\hat{\mu}(\lambda)$ and $\hat{\sigma}(\lambda)$. The estimate $\hat{\lambda}_M$ will be the value of λ which maximizes $L(\hat{\mu}(\lambda), \hat{\sigma}(\lambda), \lambda)$. We will refer to this method as the M method in the following.
The likelihood ratio statistic under model (2.3) for testing an hypothesis of the form $H_0: \quad \lambda = \lambda_0$ is given by

$$\Lambda_M = -2\ln\left(\frac{L(\hat{\mu}(\lambda_0), \hat{\sigma}(\lambda_0), \lambda_0)}{L(\hat{\mu}(\lambda_M), \hat{\sigma}(\lambda_M), \lambda_M)}\right) \tag{2.21}$$

which Carroll (1980) claims, under appropriate conditions, to be asymptotically χ^2 distributed with one degree of freedom if H_0 is true.

In an independent paper, Bickel & Doksum (1981) also suggested a similar estimator and likelihood ratio test statistic. Their theoretical calculations and simulations led to the same conclusions as in Carroll (1980), namely that his robust likelihood ratio statistic Λ_M is preferable to the method of Atkinson (1973) and the maximum-likelihood method in the sense that they all have comparable power, but the type I error of the former is much closer to the nominal level $\alpha = 0.05$, when testing the null hypothesis $H_0 : \lambda = -1$ with the true transformation parameter $\lambda = -1$ and considering a normal, two contaminated normal and a Student-t distributed error model, respectively. Even though the M method exhibits a slight increase in level considering distributions with heavy tails, it is still preferable to the significance method of Andrews (1971) because the M method is more robust.

Due to a suggestion of Bickel & Doksum (1981), Carroll (1982) modifies his estimation procedure and replaced equation (2.20) by

$$\sum_{i=1}^{n} \left[r_i(\lambda) \psi \left(r_i(\lambda) \right) - 1 \right] = 0 \, , \tag{2.22}$$

where

$$r_i(\lambda) = \frac{y_i^{(\lambda)} - \mu}{\sigma} \, .$$

Additionally, the test statistic Λ_M is replaced by

$$\Lambda_M^* = \Lambda_M \frac{\sum_{i=1}^{n} \psi' \left(r_i(\lambda) \right)}{\sum_{i=1}^{n} \psi^2 \left(r_i(\lambda) \right)} . \tag{2.23}$$

Only this improved version will be considered in the future comparisons.

2.3.2 Robust Shapiro-Wilk estimator of λ

Due to the fact that the Shapiro-Wilk test is one of the most powerful tests of normality and the performances of a non-robust estimator of λ based on it (see Gaudard & Karson (2000)), we use our robust asymmetric Shapiro-Wilk test from Chapter 1 to estimate λ

robustly. This means that we calculate the value of λ between 0 and 1 which maximizes the test statistic of the robust Shapiro-Wilk test, so that our robust Shapiro-Wilk estimate is given by

$$\lambda_{RSW} = \arg\max_{\lambda \in [0,1]} \frac{\left(\sum_{i=1}^{n} a_i y_i^{(\lambda)}\right)^2}{\sum_{i=1}^{n}\left(y_i^{(\lambda)} - \overline{y^{(\lambda)}}\right)^2}.$$

2.4 Simulations

In this section, we compare the robust and the maximum-likelihood estimators because we are solely interested in estimation in the presence of outliers. The estimators we consider are:

- The maximum likelihood estimate of λ : $\hat{\lambda}_{ML}$
- The M estimate of λ proposed by Carroll (1982): $\hat{\lambda}_M$
- The estimate of λ based on the robust Shapiro-Wilk statistic : $\hat{\lambda}_{RSW}$

All estimators are computed in the interval [0,1].

We simulate samples of length 100 from a normal distribution with mean 7 and standard-deviation 1 as follows:

- 1000 samples without outliers
- 1000 samples with one positive outlier
- 1000 samples with one positive and one negative outlier
- 1000 samples with two positive outliers

All outliers are of magnitude 6 standard deviations. Then we transform the data with the inverse Box-Cox transformation with true parameter $\lambda = 0$ (lognormal distribution), $\lambda = 0.4$ or $\lambda = 1$ (normal distribution).

To compare the different robust estimators, we compute the bias as the mean of the deviations from λ, the mean squared error (MSE) and the variance of the estimators. Since we are mainly interested in how close the transformed data are to the normal distribution, we additionally consider the p values of the robust Shapiro-Wilk test applied to the transformed samples with the Box-Cox transformation using the estimated parameter. Because $\hat{\lambda}_{RSW}$ is likely to have an advantage in such a comparison due to its definition, we also compute the medcouple to assess the symmetry of the transformed data and apply one of the robust tests based on the medcouple $MC1$ as robust measure of normality for the transformed data.

2.4.1 Estimation of the transformation parameter

Figures 2.1, 2.2 and 2.3 show boxplots of the robust estimates of λ for clean and contaminated samples and $\lambda = 0$, $\lambda = 0.4$ and $\lambda = 1$, respectively. When there are no outliers the robust methods and the maximum likelihood estimator behave similarly and are almost unbiased. In the presence of one or two positive outliers, all estimators are biased towards zero, but $\hat{\lambda}_{RSW}$ is the least affected. Interestingly, in the case of one positive and one negative outlier all estimators seem to be little affected. This can be explained by the fact that the Box-Cox transformation is meant to transform to approximate normality and hence the estimators are not or very little affected by symmetric configurations of the outliers.

The biases of the estimators are recapitulated in Table 2.1. We observe that all estimators are negatively biased except for the case where the true transformation parameter is $\lambda = 0$, because we have restricted the estimators to be positive. We notice that $\hat{\lambda}_{RSW}$ has the smallest bias except for $\lambda = 0$ and it is more robust in comparison to the M estimates. Only in the case of clean samples is $\hat{\lambda}_{RSW}$ outperformed by $\hat{\lambda}_{ML}$, but it always behaves rather well. We see that all other estimators are biased towards 0 in the presence of outliers, except if there is one positive and one negative outlier, as mentioned before. This can be explained by the fact that in the presence of only positive outliers, a Box-Cox transformation with a small value of the transformation parameter would attempt to eliminate the skewness in the data due to the outliers. Our estimator $\hat{\lambda}_{RSW}$ takes into account the skewness in the data and is hence less affected by the outliers.

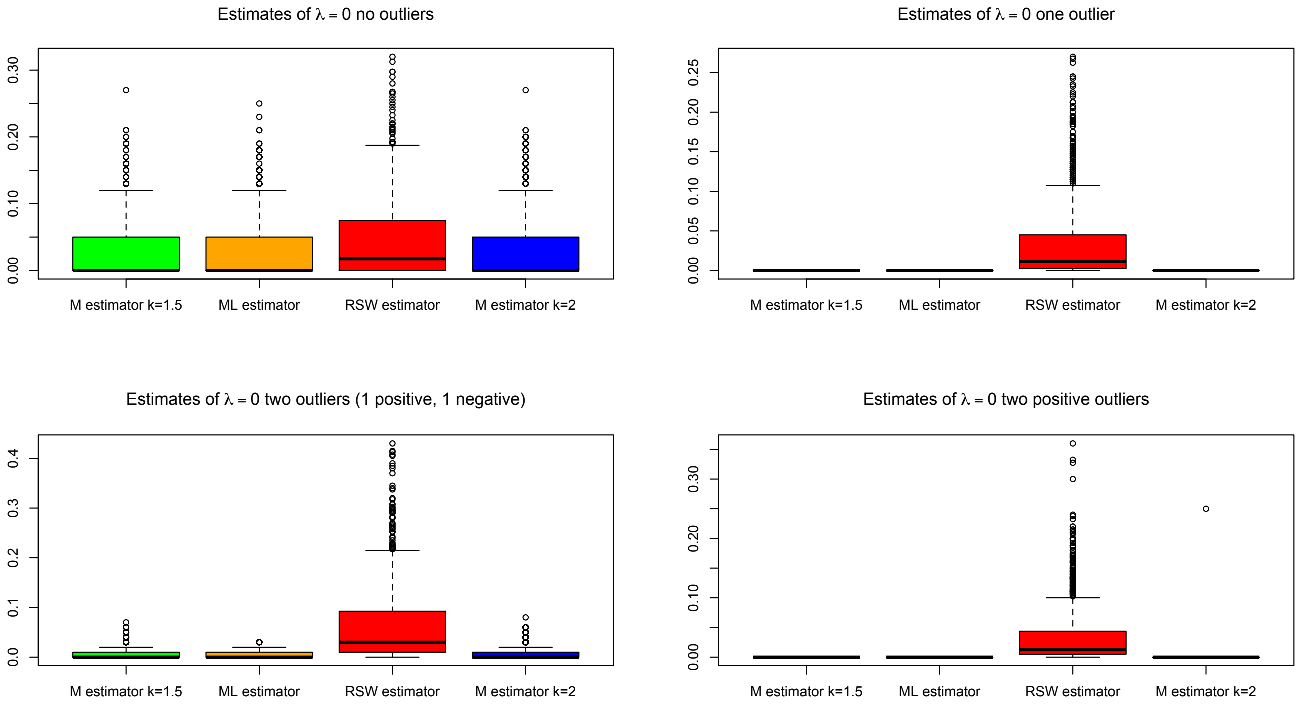
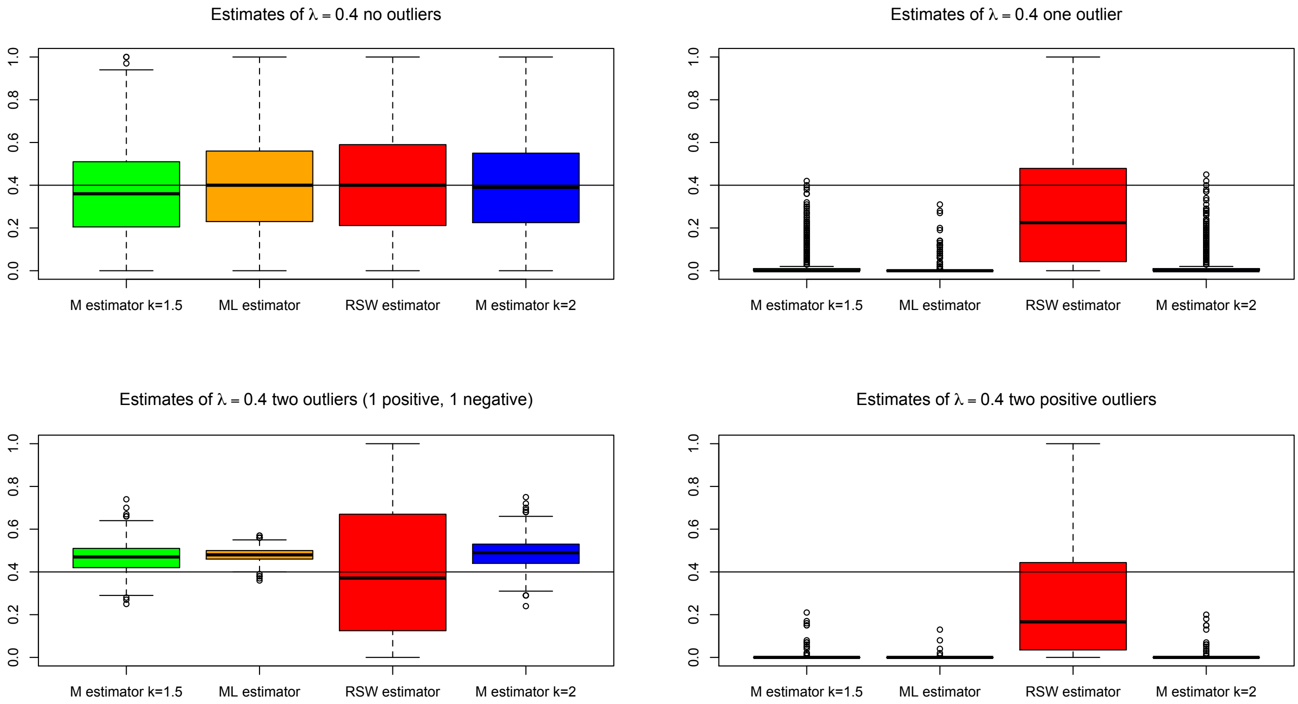


Figure 2.1: Boxplot of the estimate of λ for the different scenarios when the true transformation parameter is $\lambda = 0$.

We compute the standard deviations of the estimators, to assess the variability of the estimators. Table 2.3 reveals that $\hat{\lambda}_{RSW}$ is more volatile than the other estimators, while $\hat{\lambda}_{ML}$ has the smallest variance. The M-estimates outperform $\hat{\lambda}_{RSW}$ in terms of variability. Table 2.2 confirms the conclusions that were drawn from Table 2.1. The $\hat{\lambda}_{RSW}$ outperforms the other estimates in the presence of outliers also in terms of mean squared error, except for $\lambda = 0$. Apparently, the MSE is dominated by the bias in the presence of outliers.



Estimates of λ = 0.4 no outliers

Estimates of λ = 0.4 one outlier

Estimates of λ = 0.4 two outliers (1 positive, 1 negative)

Estimates of λ = 0.4 two positive outliers

Figure 2.2: Boxplot of the estimate of λ for the different scenarios when the true transformation parameter is $\lambda = 0.4$.

2.4.2 Comparison of the estimators in terms of best transformation

Because we are mainly interested in the parameter estimate that yields transformed samples which are closest to the normal distribution, we compare the estimators according to criteria based on this idea, applying our robust asymmetric Shapiro-Wilk test and the $MC1$ test to the transformed samples. We also compute the medcouple of the transformed samples as a robust measure of symmetry.

2.4.2.1 Power of our robust Shapiro-Wilk test

As we expected, Figures 2.4, 2.5 and 2.6 show that the p values of the robust Shapiro-Wilk test applied to the samples transformed with $\hat{\lambda}_{RSW}$ are typically larger than those

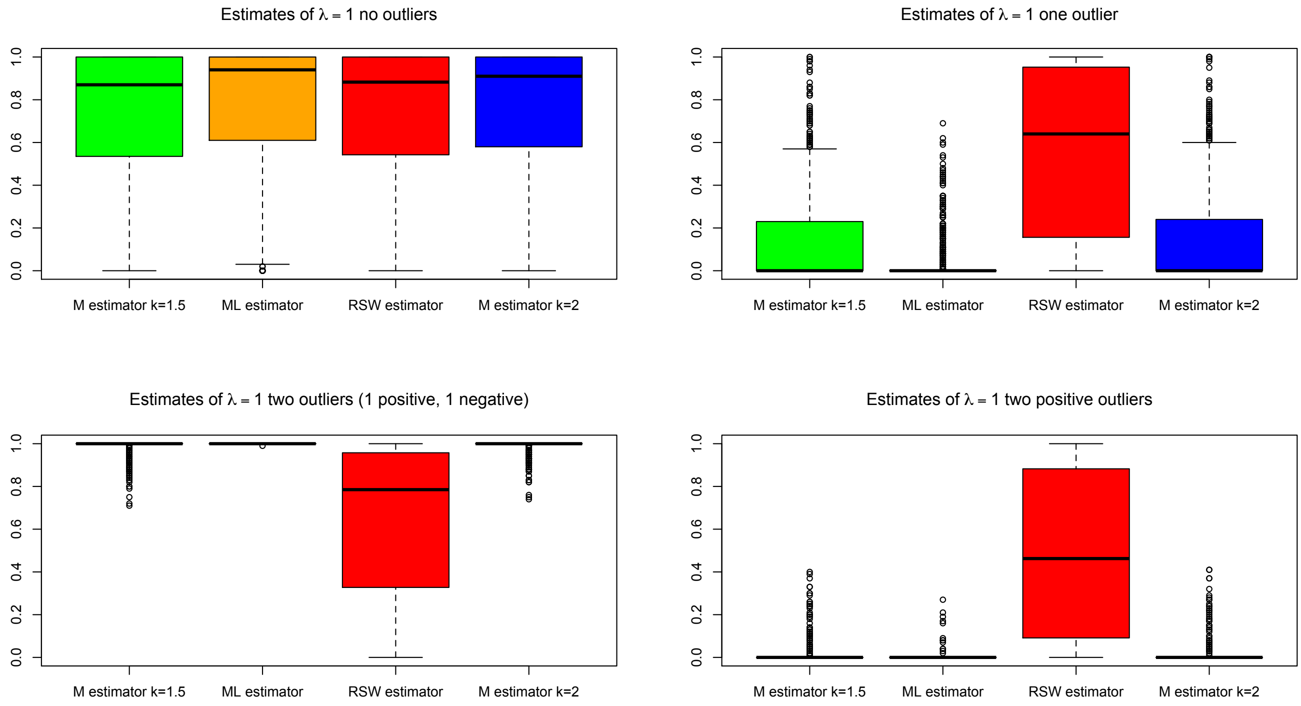


Figure 2.3: Boxplot of the estimate of λ for the different scenarios when the true transformation parameter is $\lambda = 1$.

obtained by the other estimators, except when the true parameter is $\lambda = 0$. Table 2.4 lists the pass rates of the robust Shapiro-Wilk test for the transformed sequences with the different parameter estimates at a significance level of 5% and confirms our previous conclusion. We note that all estimators perform quite well in this respect, surprisingly even the maximum likelihood estimate that yields very large pass rates. This can be explained by the fact that in the presence of outliers, the maximum likelihood estimate is biased towards 0 and transforming with $\lambda = 0$ seems to yield good results independently of the true value of the transformation parameter.

2.4.2.2 Power of $MC1$ test

The results in Figures 2.7, 2.8 and 2.9 show the p values of the $MC1$ test applied to the samples transformed with the different robust parameter estimates and Table 2.5

$\lambda = 0$				
Scenario	$\hat{\lambda}_{ML}$	$\hat{\lambda}_{RSW}$	$\hat{\lambda}_M$ with k=1.5	$\hat{\lambda}_M$ with k=2
no outlier	0.033	0.046	0.032	0.033
one outlier	0.000	0.034	0.000	0.000
one positive and one negative outlier	0.004	0.063	0.008	0.008
two positive outliers	0.000	0.034	0.000	0.000
$\lambda = 0.4$				
Scenario	$\hat{\lambda}_{ML}$	$\hat{\lambda}_{RSW}$	$\hat{\lambda}_M$ with k=1.5	$\hat{\lambda}_M$ with k=2
no outlier	-0.003	0.008	-0.036	-0.011
one outlier	-0.396	-0.109	-0.371	-0.371
one positive and one negative outlier	0.079	0.015	0.069	0.085
two positive outliers	-0.400	-0.135	-0.398	-0.399
$\lambda = 1$				
Scenario	$\hat{\lambda}_{ML}$	$\hat{\lambda}_{RSW}$	$\hat{\lambda}_M$ with k=1.5	$\hat{\lambda}_M$ with k=2
no outlier	-0.225	-0.265	-0.261	-0.237
1 outlier	-0.973	-0.439	-0.864	-0.861
two outliers (positive,negative)	0.000	-0.360	-0.009	-0.006
2 postive outliers	-0.999	-0.519	-0.991	-0.992

Table 2.1: Biases of the the maximum likelihood and the robust estimators.

contains the pass rates for the $MC1$ test at a significance level of 5%. All estimators perform similarly, with slightly better results for $\hat{\lambda}_{RSW}$ than for the other estimators, if the true transformation parameter is different from 0. It is interesting to note that even though sometimes the estimators of λ are biased, the $MC1$ test asserts that the transformed samples with the respective estimators are mostly normal and $\hat{\lambda}_{RSW}$ yields slightly better results than the others except in the case $\lambda = 0$, but the differences are negligible.

$\lambda = 0$				
Scenario	$\hat{\lambda}_{ML}$	$\hat{\lambda}_{RSW}$	$\hat{\lambda}_M$ with k=1.5	$\hat{\lambda}_M$ with k=2
no outlier	0.059	0.076	0.056	0.059
one outlier	0.000	0.060	0.000	0.000
one positive and one negative outlier	0.007	0.098	0.016	0.016
two positive outliers	0.000	0.060	0.000	0.008
$\lambda = 0.4$				
Scenario	$\hat{\lambda}_{ML}$	$\hat{\lambda}_{RSW}$	$\hat{\lambda}_M$ with k=1.5	$\hat{\lambda}_M$ with k=2
no outlier	0.227	0.252	0.217	0.225
one outlier	0.397	0.293	0.377	0.377
one positive and one negative outlier	0.086	0.313	0.098	0.111
two positive outliers	0.400	0.298	0.399	0.399
$\lambda = 1$				
Scenario	$\hat{\lambda}_{ML}$	$\hat{\lambda}_{RSW}$	$\hat{\lambda}_M$ with k=1.5	$\hat{\lambda}_M$ with k=2
no outlier	0.369	0.412	0.399	0.379
one outlier	0.977	0.580	0.890	0.889
one positive and one negative outlier	0.000	0.501	0.033	0.026
two positive outliers	0.999	0.636	0.992	0.993

Table 2.2: Root of the mean squared errors of the robust and the maximum likelihood estimators.

2.4.2.3 Medcouple of the transformed samples

When a sample is exactly symmetric its medcouple is 0, so that departures from symmetry can be measured by the absolute value of the medcouple. Table 2.6 shows that the robust estimators yield more symmetric samples after transformation than the maximum-likelihood estimator and that $\hat{\lambda}_{RSW}$ has the best results in all cases even when the true transformation parameter is $\lambda = 0$. But again we should stress the fact that the departures from symmetry of the transformed samples with $\hat{\lambda}_{ML}$ are not very large with respect to

$\lambda = 0$				
Scenario	$\hat{\lambda}_{ML}$	$\hat{\lambda}_{RSW}$	$\hat{\lambda}_{M}$ with k=1.5	$\hat{\lambda}_{M}$ with k=2
no outlier	0.048	0.060	0.046	0.048
one outlier	0.000	0.049	0.000	0.000
one positive and one negative outlier	0.006	0.076	0.013	0.014
two positive outliers	0.000	0.050	0.000	0.008
$\lambda = 0.4$				
Scenario	$\hat{\lambda}_{ML}$	$\hat{\lambda}_{RSW}$	$\hat{\lambda}_{M}$ with k=1.5	$\hat{\lambda}_{M}$ with k=2
no outlier	0.227	0.252	0.214	0.225
one outlier	0.024	0.272	0.066	0.067
one positive and one negative outlier	0.032	0.312	0.069	0.071
two positive outliers	0.005	0.266	0.013	0.013
$\lambda = 1$				
Scenario	$\hat{\lambda}_{ML}$	$\hat{\lambda}_{RSW}$	$\hat{\lambda}_{M}$ with k=1.5	$\hat{\lambda}_{M}$ with k=2
no outlier	0.293	0.316	0.302	0.296
one outlier	0.087	0.379	0.217	0.221
one positive and one negative outlier	0.000	0.349	0.032	0.026
two positive outliers	0.015	0.369	0.044	0.042

Table 2.3: Standard deviations of the robust and the maximum likelihood estimators.

the others.

2.5 Conclusion

Due to the fact that the assumption of normality is very common and that in practice, data are usually not normally distributed, extensive work has been done to solve this problem using transformations. Many authors have studied the popular Box-Cox transformation and estimators have been developed over the years. Because of outliers the need

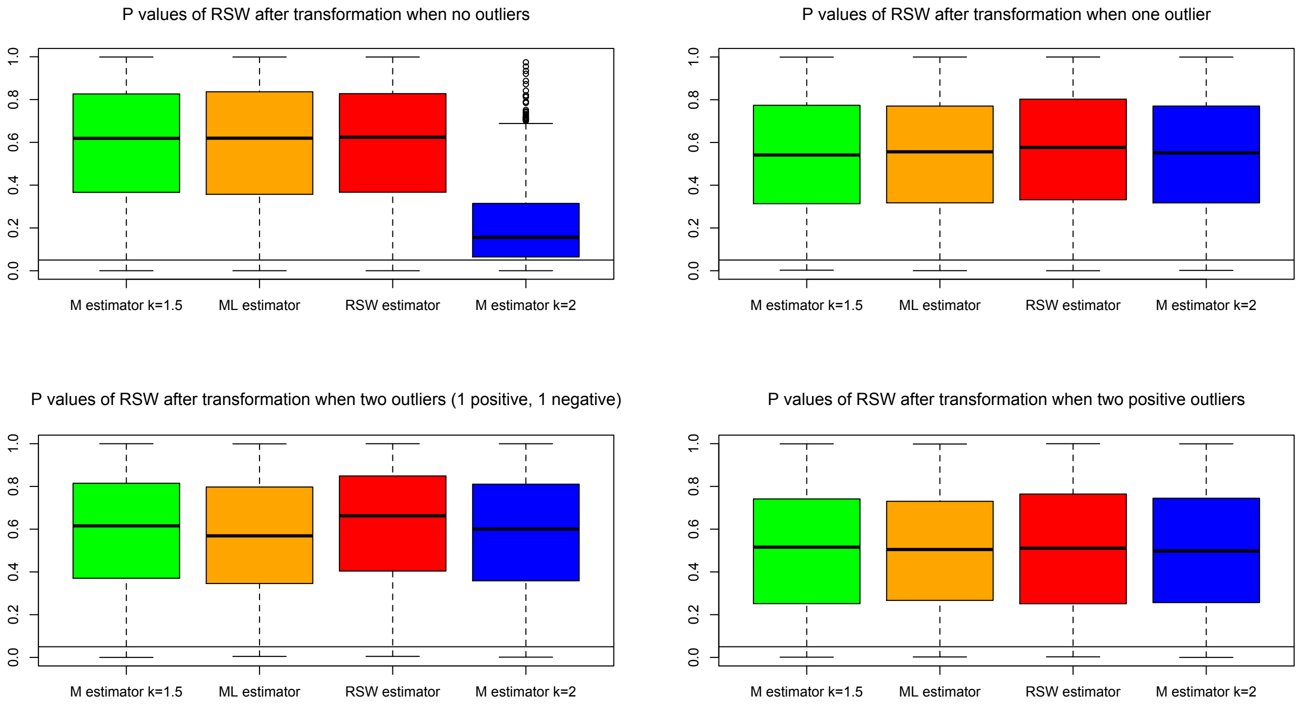


Figure 2.4: Boxplot of p values of our robust Shapiro-Wilk test applied to the transformed samples when the true transformation parameter is $\lambda = 0$.

for robust estimates of the transformation parameter λ has increased. Carroll (1982) has developed an M-estimate of the transformation parameter ($\hat{\lambda}_M$) and illustrated its advantages to outperform the maximum-likelihood estimator ($\hat{\lambda}_{ML}$).

We introduce a robust estimator ($\hat{\lambda}_{RSW}$) of the transformation parameter based on the maximization of our robust Shapiro-Wilk test statistic. Simulations show that it outperforms the M-estimator and the maximum-likelihood estimator in various cases, with the later not being robust against outliers. Even though $\hat{\lambda}_{RSW}$ has a larger variance than $\hat{\lambda}_M$ and $\hat{\lambda}_{ML}$, it is less biased in the presence of outliers of the same sign and its mean squared error is lower than those of the others.

Furthermore, keeping in mind that our main goal is to provide the closest transformation to normality, we also provide evidence via simulations that transforming with $\hat{\lambda}_{RSW}$ yields more symmetrical samples, which are generally closer to normality in the sense of higher

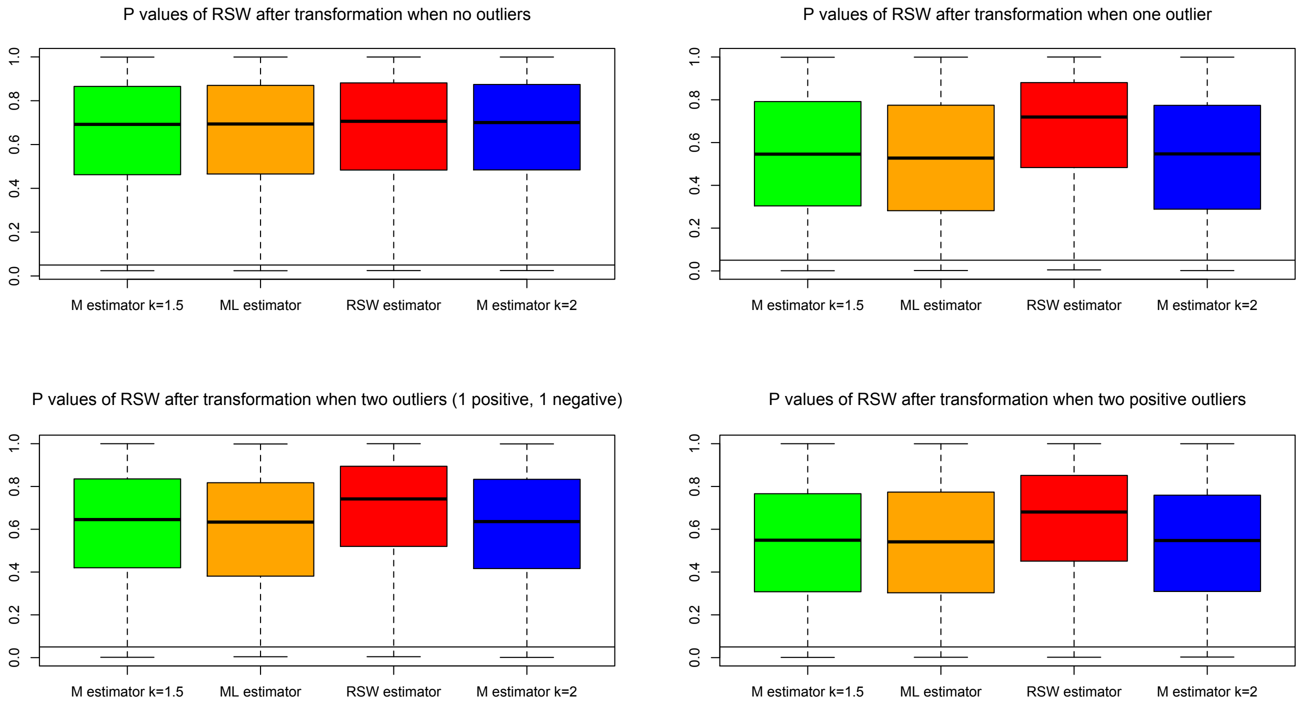


Figure 2.5: Boxplot of p values of our robust Shapiro-Wilk test applied to the transformed samples when the true transformation parameter is $\lambda = 0.4$.

p values of our robust Shapiro-Wilk test and a robust test based on the medcouple, which was used as a robust measure of symmetry for the transformed samples.

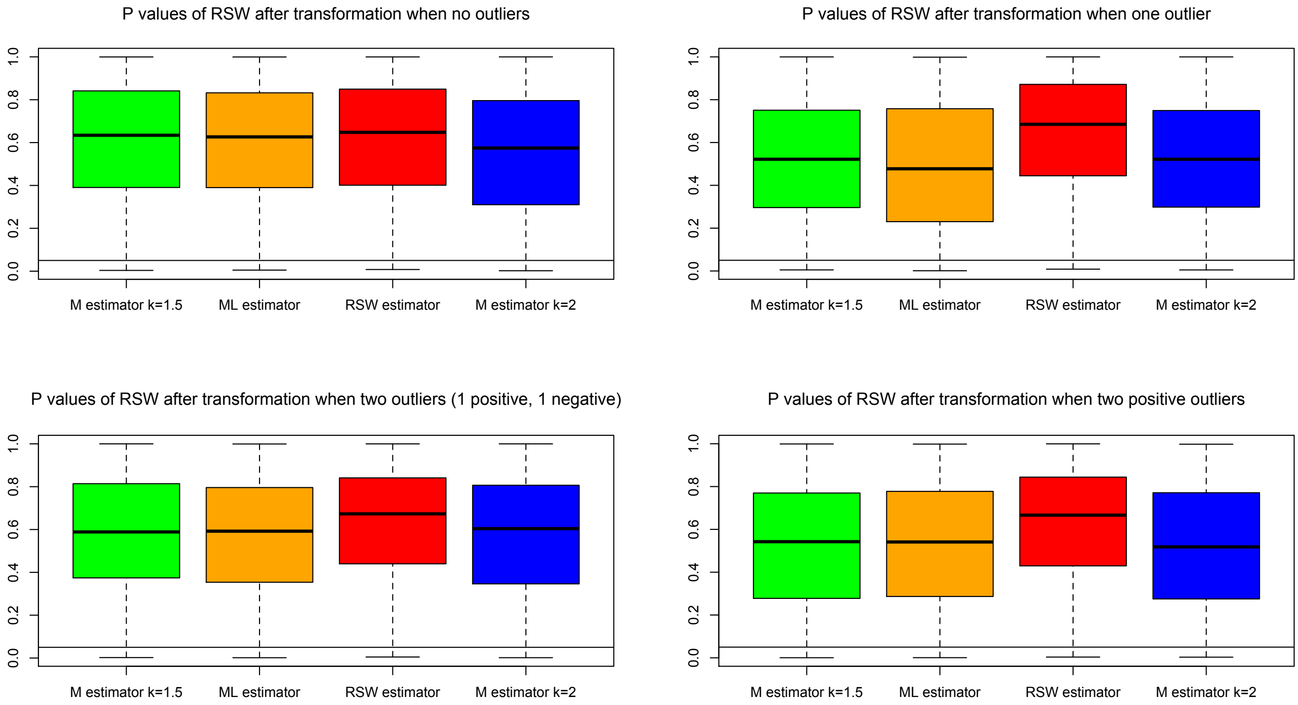
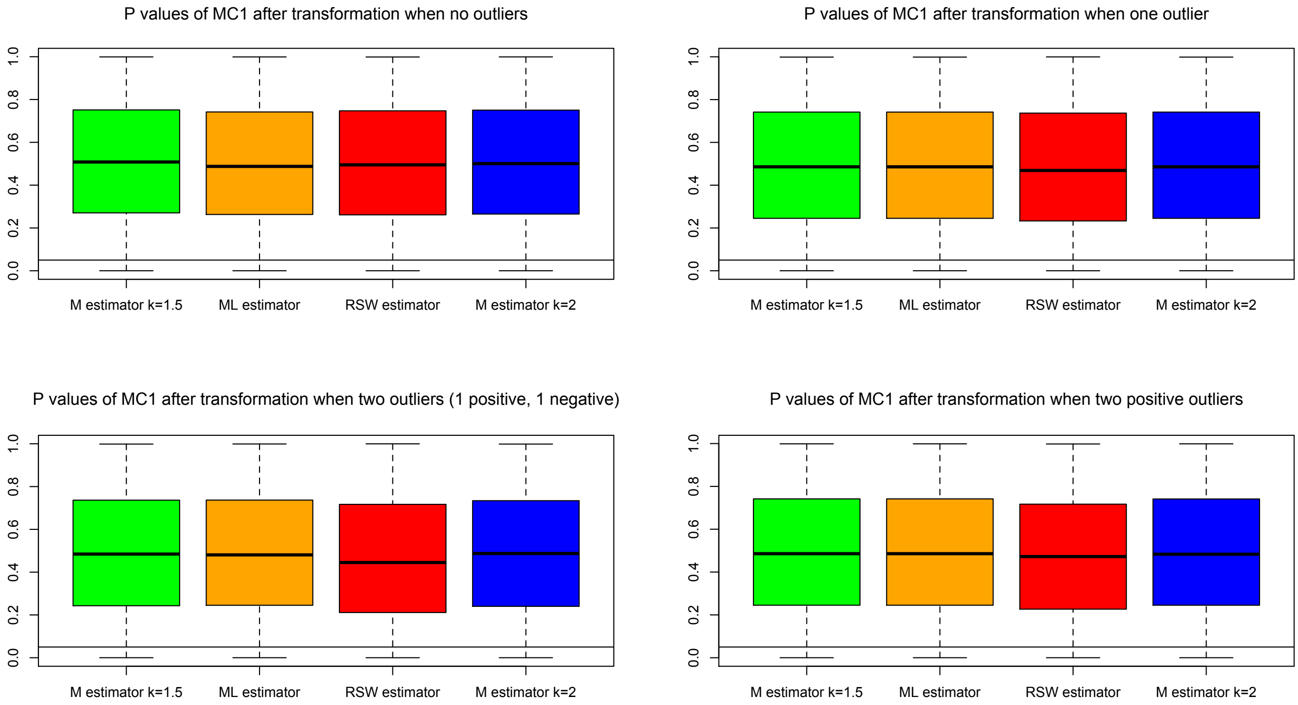


Figure 2.6: Boxplot of p values of our robust Shapiro-Wilk test applied to the transformed samples when the true transformation parameter is $\lambda = 1$.

$\lambda = 0$				
Scenario	$\hat{\lambda}_{ML}$	$\hat{\lambda}_{RSW}$	$\hat{\lambda}_{M}$ with k=1.5	$\hat{\lambda}_{M}$ with k=2
no outlier	97.6	98.3	98.3	79.2
one outlier	96.6	97.5	96.4	96.7
one positive and one negative outlier	98.2	98.9	98.7	98.2
two positive outliers	94.1	94.3	94.5	95.0
$\lambda = 0.4$				
Scenario	$\hat{\lambda}_{ML}$	$\hat{\lambda}_{RSW}$	$\hat{\lambda}_{M}$ with k=1.5	$\hat{\lambda}_{M}$ with k=2
no outlier	99.4	99.5	99.3	99.4
one outlier	94.2	98.8	97.6	97.2
one positive and one negative outlier	98.5	99.4	98.5	98.3
two positive outliers	95.8	97.9	96.3	95.9
$\lambda = 1$				
Scenario	$\hat{\lambda}_{ML}$	$\hat{\lambda}_{RSW}$	$\hat{\lambda}_{M}$ with k=1.5	$\hat{\lambda}_{M}$ with k=2
no outlier	98.2	98.1	98.2	97.3
one outlier	93.8	98.8	97.9	97.9
one positive and one negative outlier	97.9	98.3	97.7	97.9
two positive outliers	96.2	98.9	96.6	97.0

Table 2.4: Pass rates for our robust Shapiro-Wilk test at a significance level of 0.05 applied to the transformed samples with the robust and maximum likelihood estimates of the transformation parameter in percent(%).



P values of MC1 after transformation when no outliers

P values of MC1 after transformation when one outlier

P values of MC1 after transformation when two outliers (1 positive, 1 negative)

P values of MC1 after transformation when two positive outliers

Figure 2.7: Boxplot of p values of the $MC1$ test applied to the transformed samples when the true transformation parameter is $\lambda = 0$.

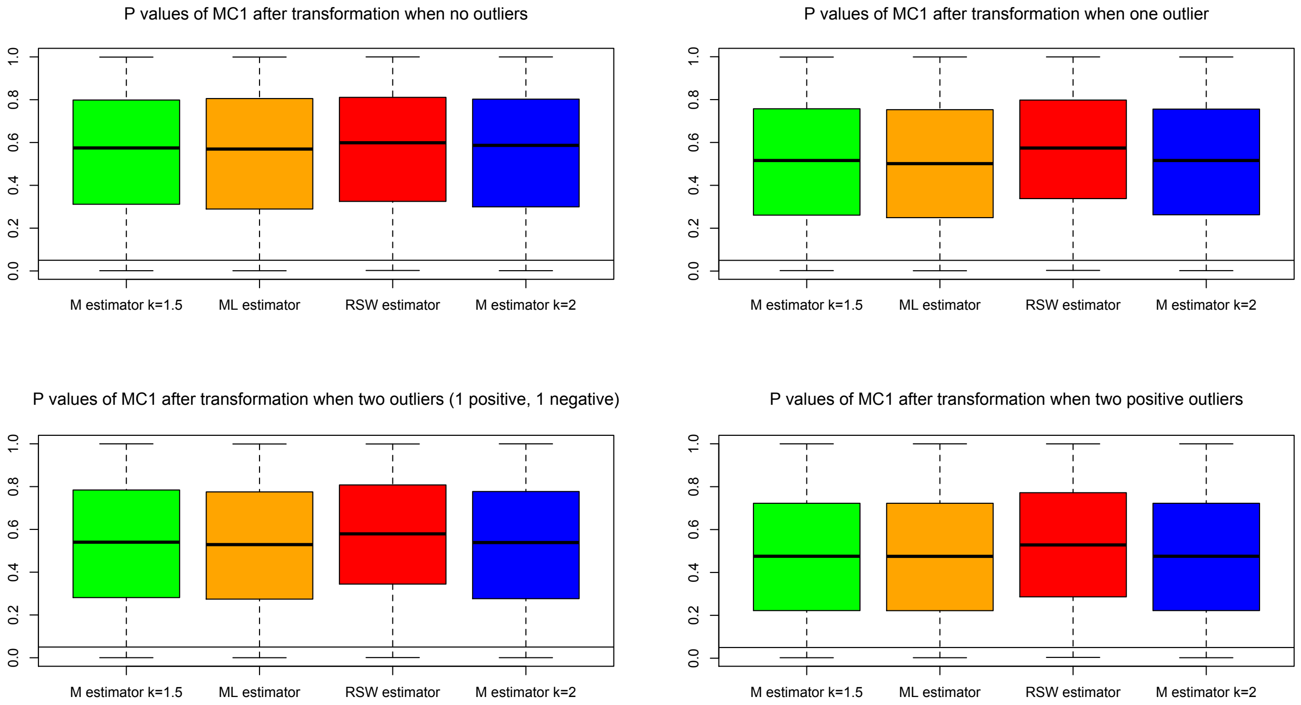


Figure 2.8: Boxplot of p values of the $MC1$ test applied to the transformed samples when the true transformation parameter is $\lambda = 0.4$.

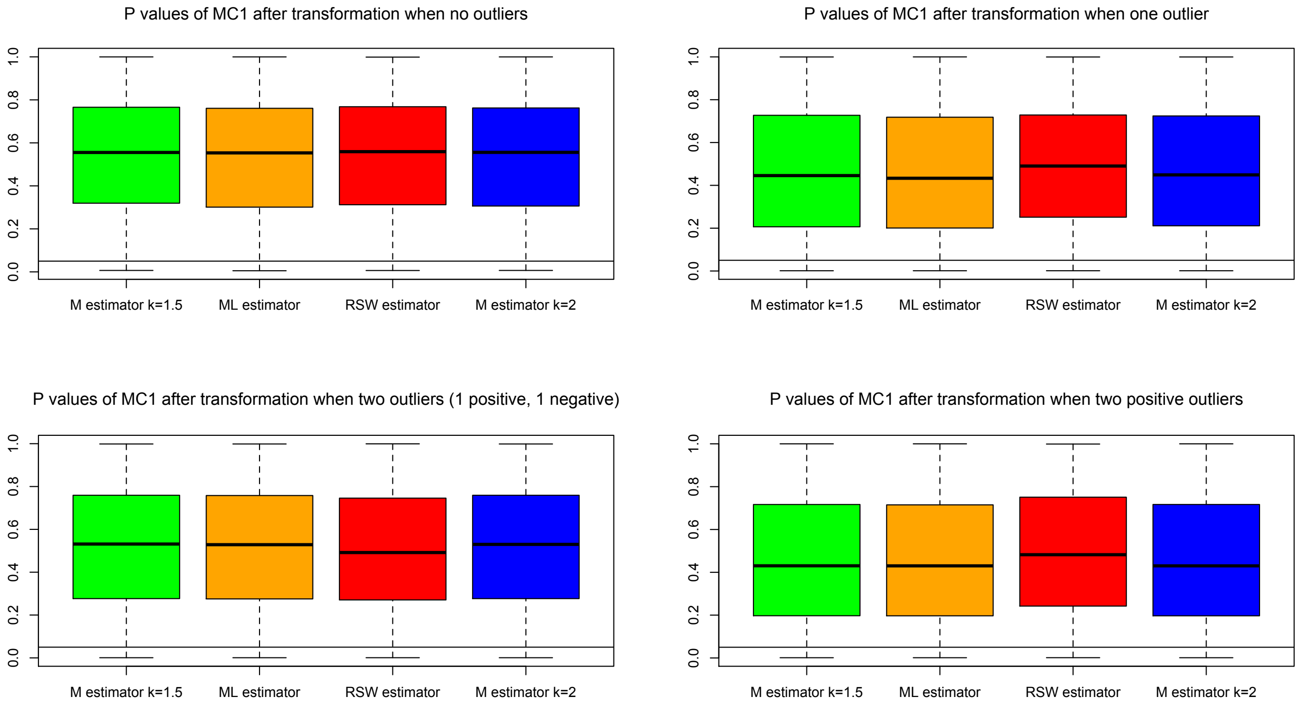


Figure 2.9: Boxplot of p values of the $MC1$ test applied to the transformed samples when the true transformation parameter is $\lambda = 1$.

$\lambda = 0$				
Scenario	$\hat{\lambda}_{ML}$	$\hat{\lambda}_{RSW}$	$\hat{\lambda}_M$ with k=1.5	$\hat{\lambda}_M$ with k=2
no outlier	95.6	96.0	96.2	95.9
one outlier	94.7	94.4	94.7	94.7
one positive and one negative outlier	94.7	93.3	94.5	94.6
two positive outliers	94.7	94.2	94.7	94.7
$\lambda = 0.4$				
Scenario	$\hat{\lambda}_{ML}$	$\hat{\lambda}_{RSW}$	$\hat{\lambda}_M$ with k=1.5	$\hat{\lambda}_M$ with k=2
no outlier	96.3	96.9	97.2	96.5
one outlier	94.3	97.8	94.8	94.8
one positive and one negative outlier	96.4	97.7	96.5	96.5
two positive outliers	95.8	97.0	95.8	95.8
$\lambda = 1$				
Scenario	$\hat{\lambda}_{ML}$	$\hat{\lambda}_{RSW}$	$\hat{\lambda}_M$ with k=1.5	$\hat{\lambda}_M$ with k=2
no outlier	96.2	96.2	96.1	96.2
one outlier	91.4	93.8	91.9	92.0
one positive and one negative outlier	95.4	94.4	95.4	95.4
two positive outliers	91.2	93.6	91.2	91.2

Table 2.5: Pass rates of the $MC1$ test at a significance level of 0.05 applied to the transformed samples with the robust and maximum likelihood estimates of the transformation parameter in percent(%).

<table>
<tr><td colspan="5">$\lambda = 0$</td></tr>
<tr><td>Scenario</td><td>$\hat{\lambda}_{ML}$</td><td>$\hat{\lambda}_{RSW}$</td><td>$\hat{\lambda}_{M}$ with k=1.5</td><td>$\hat{\lambda}_{M}$ with k=2</td></tr>
<tr><td>no outlier</td><td>0.089</td><td>0.088</td><td>0.087</td><td>0.088</td></tr>
<tr><td>one outlier</td><td>0.089</td><td>0.085</td><td>0.089</td><td>0.089</td></tr>
<tr><td>one positive and one negative outlier</td><td>0.090</td><td>0.089</td><td>0.090</td><td>0.090</td></tr>
<tr><td>two positive outliers</td><td>0.094</td><td>0.093</td><td>0.094</td><td>0.094</td></tr>
<tr><td colspan="5">$\lambda = 0.4$</td></tr>
<tr><td>Scenario</td><td>$\hat{\lambda}_{ML}$</td><td>$\hat{\lambda}_{RSW}$</td><td>$\hat{\lambda}_{M}$ with k=1.5</td><td>$\hat{\lambda}_{M}$ with k=2</td></tr>
<tr><td>no outlier</td><td>0.080</td><td>0.075</td><td>0.077</td><td>0.078</td></tr>
<tr><td>one outlier</td><td>0.089</td><td>0.076</td><td>0.087</td><td>0.087</td></tr>
<tr><td>one positive and one negative outlier</td><td>0.084</td><td>0.074</td><td>0.082</td><td>0.082</td></tr>
<tr><td>two positive outliers</td><td>0.093</td><td>0.082</td><td>0.093</td><td>0.093</td></tr>
<tr><td colspan="5">$\lambda = 1$</td></tr>
<tr><td>Scenario</td><td>$\hat{\lambda}_{ML}$</td><td>$\hat{\lambda}_{RSW}$</td><td>$\hat{\lambda}_{M}$ with k=1.5</td><td>$\hat{\lambda}_{M}$ with k=2</td></tr>
<tr><td>no outlier</td><td>0.081</td><td>0.079</td><td>0.080</td><td>0.080</td></tr>
<tr><td>one outlier</td><td>0.092</td><td>0.080</td><td>0.088</td><td>0.088</td></tr>
<tr><td>one positive and one negative outlier</td><td>0.090</td><td>0.085</td><td>0.089</td><td>0.090</td></tr>
<tr><td>two positive outliers</td><td>0.091</td><td>0.080</td><td>0.090</td><td>0.091</td></tr>
</table>

Table 2.6: Medcouples of the transformed samples with the robust and maximum likelihood estimates of the transformation parameter.

Chapter 3

Robust online state change detection in time series

3.1 Introduction

Monitoring time series has become a very challenging task over the years due to the increasing amount of complexity introduced in statistical models. State change detection gives useful insight on the behaviour of a process and allows a deeper understanding and a better management of the monitored system. A state change, simply put, is a change of behaviour in a process. One of the most common state changes is the outlier, which can be defined as an unusual observation lying far from the bulk of the majority of the data. For example, if we are monitoring the number of cars produced in a factory per day, an outlier can be caused by a strike of employees for a day, in which case few to no cars will be produced on that given day. Our aim in this chapter is to build a powerful state change detection procedure to monitor changes in time series. In Section 2, we present the Harrison and Stevens method to model state changes by assuming a state space model with a local linear trend and standard normal errors. The estimation procedure is a Bayesian method coupled to the Kalman Filter and in Section 3, we rectify an error in the Kalman Filter used to update the model parameters. Then we extend the corrected procedure in Section 4. Some improvement to the procedure are presented in Section 5 and Section 6 deals with an example of application of the enhanced procedure. Section 7 is a short conclusion.

3.2 The Harrison & Stevens method

Harrison & Stevens (1971) developed a Bayesian method for analysing time series with outliers and sudden changes based on a local linear trend model of the form

$$\begin{aligned} y_t &= \mu_t s_t + \epsilon_t \\ \mu_t &= \mu_{t-1} + \beta_{t-1} + \gamma_t \\ \beta_t &= \beta_{t-1} + \delta_t, \end{aligned} \tag{3.1}$$

where ϵ_t, γ_t and δ_t are independent normally distributed random errors with mean 0 and variance V_ϵ, V_γ and V_δ, respectively, and s_t is a seasonal factor.

The model are fitted within a Bayesian framework along with the probability of occurrence of one of several states like outliers, . . . , given all previous observations. For simplification, we neglect the seasonal factor s_t and set it to one.

Before going further into details, we present the different states that are to be modelled. In our case, we consider four states: steady state, step change, slope change and transient (also known as outlier). If the system is in a state j at time t the random components ϵ_t, γ_t and δ_t are assumed to be generated by normal distributions with mean 0 and variance $V_\epsilon^{(j)}$, $V_\gamma^{(j)}$ and $V_\delta^{(j)}$, $j = 1, 2, 3$ and 4, respectively, for each of the four states. The four states are defined as

- steady state, if $V_\epsilon^{(j)}$ is normal, $V_\gamma^{(j)}$ and $V_\delta^{(j)}$ are null
- step change, if $V_\epsilon^{(j)}$ is normal, $V_\gamma^{(j)}$ is large and $V_\delta^{(j)}$ is null
- slope change, if $V_\epsilon^{(j)}$ is normal, $V_\gamma^{(j)}$ is null and $V_\delta^{(j)}$ is large
- outlier, if $V_\epsilon^{(j)}$ is large, $V_\gamma^{(j)}$ and $V_\delta^{(j)}$ are null

In Figure 3.1 we can see simulated data of length 200 in which we have incorporated a state change at time 100.

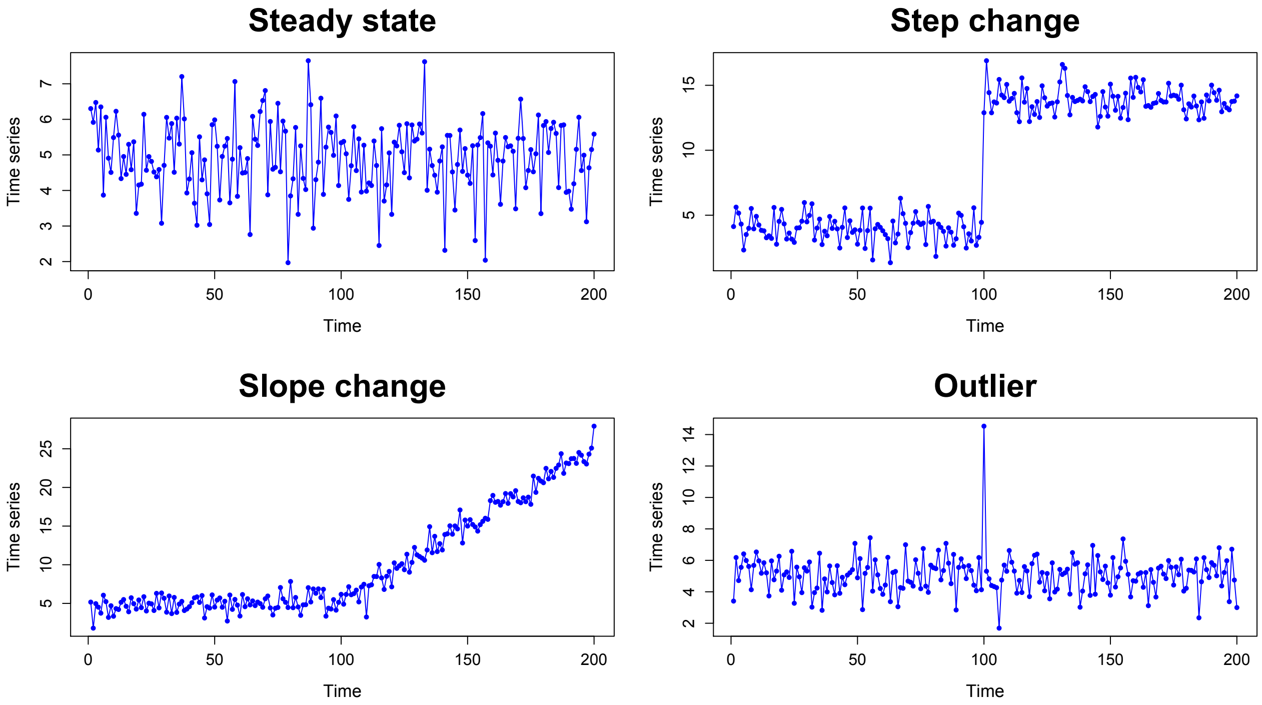


Figure 3.1: Four states considered in Harrison & Stevens (1971).

Define $\psi = \{m, b, v_{\mu\mu}, v_{\mu\beta}, v_{\beta\beta}\}$, where

$$\begin{aligned} m &= E(\mu) \\ b &= E(\beta) \\ v_{\mu\mu} &= E(\mu - m)^2 \\ v_{\mu\beta} &= E\left[(\mu - m)(\beta - b)\right] \\ v_{\beta\beta} &= E(\beta - b)^2 \,. \end{aligned}$$

Suffices and subscripts on ψ are to be associated with all its components. To indicate that any pair of variables (μ, β) is jointly bivariate normally distributed with parameter ψ, we write $(\mu, \beta) \sim N(\psi)$.

Harrison & Stevens (1971) show that if the joint distribution of (μ, β) at time $t - 1$ is bivariate normal,

$$(\mu_{t-1}, \beta_{t-1} \mid y_{t-1}) \sim N(\psi_{t-1}) \,,$$

then the posterior distribution at time t is also bivariate normal:

$$(\mu_t, \beta_t \mid y_t) \sim N(\psi_t).$$

Harrison & Stevens used a Kalman filter to update the components of the parameter ψ_t, but there are some small mistakes in the update equations that they proposed. In the following, we will derive the correct expressions.

If we define

$$\alpha_t = \begin{pmatrix} \mu_t \\ \beta_t \end{pmatrix} \quad \text{and} \quad \eta_t = \begin{pmatrix} \gamma_t \\ \delta_t \end{pmatrix},$$

then we can redefine the state space model (3.1) as follows

$$\begin{aligned} y_t &= Z_t \alpha_t + \varepsilon_t \\ \alpha_t &= T_t \alpha_{t-1} + \eta_t \end{aligned}$$

where

$$T_t = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad Z_t = \begin{pmatrix} 1 & 0 \end{pmatrix}.$$

Let us define

$$Q_t = Var(\eta_t) \quad \text{and} \quad P_t = Var(\alpha_t) = \begin{pmatrix} v_{\mu\mu,t} & v_{\mu\beta,t} \\ v_{\mu\beta,t} & v_{\beta\beta,t} \end{pmatrix}.$$

By applying the Kalman Filter as in Harvey (1991), we obtain the following expression for the covariance matrix of the predictions

$$\begin{aligned} P_{t|t-1} =& T_t P_{t-1} T_t' + Q_t \\ =& \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} v_{\mu\mu,t-1} & v_{\mu\beta,t-1} \\ v_{\mu\beta,t-1} & v_{\beta\beta,t-1} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} + \begin{pmatrix} V_\gamma & 0 \\ 0 & V_\delta \end{pmatrix} \\ =& \begin{pmatrix} v_{\mu\mu,t-1} + 2v_{\mu\beta,t-1} + v_{\beta\beta,t-1} + V_\gamma & v_{\mu\beta,t-1} + v_{\beta\beta,t-1} \\ v_{\mu\beta,t-1} + v_{\beta\beta,t-1} & v_{\beta\beta,t-1} + V_\delta \end{pmatrix}. \end{aligned}$$

Harrison & Stevens (1971) set

$$P_{t|t-1} = R = \begin{pmatrix} r_{11} & r_{12} \\ r_{12} & r_{22} \end{pmatrix}$$

so that the following equations given by Harrison & Stevens

$$\begin{aligned} r_{11} &= v_{\mu\mu,t-1} + 2v_{\mu\beta,t-1} + v_{\beta\beta,t-1} + V_\gamma + V_\delta \\ r_{12} &= v_{\mu\beta,t-1} + v_{\beta\beta,t-1} + V_\delta \\ r_{22} &= v_{\beta\beta,t-1} + V_\delta \end{aligned}$$

can be corrected to

$$r_{11} = v_{\mu\mu,t-1} + 2v_{\mu\beta,t-1} + v_{\beta\beta,t-1} + V_\gamma \tag{3.2}$$

$$r_{12} = v_{\mu\beta,t-1} + v_{\beta\beta,t-1} \tag{3.3}$$

$$r_{22} = v_{\beta\beta,t-1} + V_\delta \,. \tag{3.4}$$

Let

$$\begin{aligned} e_t &= y_t - m_{t-1} - b_{t-1} \\ V_e &= r_{11} + V_\epsilon \\ A_1 &= r_{11}/V_e \\ A_2 &= r_{12}/V_e \,. \end{aligned}$$

Then the parameter vector ψ_t of the posterior distribution at time t is given by:

$$\begin{cases} m_t & = m_{t-1} + b_{t-1} + A_1 e_t \\ b_t & = b_{t-1} + A_2 e_t \\ v_{\mu\mu,t} & = r_{11} - A_1^2 V_e \\ v_{\mu\beta,t} & = r_{12} - A_1 A_2 V_e \\ v_{\beta\beta,t} & = r_{22} - A_2^2 V_e \end{cases}$$

The parameter estimates for time t can thus be obtained from those for time $t-1$ by an update step, which can be summarized as follows:

$$\psi_t = B(\psi_{t-1}; V_\epsilon, V_\gamma, V_\delta) \,.$$

Let us now consider the case of our four states and assume that the distribution of (μ, β) is a mixture of bivariate normal distributions as follows:

$$(\mu_{t-1}, \beta_{t-1} \mid y_{t-1}) \sim \sum_{i=1}^{i=4} q_{t-1}^{(i)} N(\psi_{t-1}^{(i)}),$$

where $y(t-1)$ represents all observations before and including y_{t-1} and

$$q_{t-1}^{(i)} = Pr(S_{t-1} \mid y(t-1))$$

is the probability posterior to $y(t-1)$ that the process was in state i at time $t-1$. Further, $\psi_{t-1}^{(i)}$ are the parameters of the distribution arising from state i at time $t-1$.
Applying the previous result for each current state $j = 1, 2, 3$ and 4 we obtain:

$$(\mu_t, \beta_t \mid y_t, S_t = j, S_{t-1} = i) \sim N(\psi_t^{(i,j)}),$$

where $\psi_t^{(i,j)} = B(\psi_{t-1}^{(i)}; V_\epsilon^{(j)}, V_\gamma^{(j)}, V_\delta^{(j)})$.
The complete posterior distribution can therefore be written as

$$(\mu_t, \beta_t \mid y_t) \sim \sum_{i,j} p_t^{(i,j)} N(\psi_t^{(i,j)}),$$

so that it is just necessary to compute $p_t^{(i,j)} = P(S_t = j, S_{t-1} = i \mid y(t))$.
Let $k_t = 1/P(y_t \mid y(t-1))$, where $P(y_t \mid y(t-1))$ denotes the probability of observing y_t at time t given all the observations until time $t-1$. It follows:

$$\begin{aligned}
p_t^{(i,j)} &= P(S_t = j, S_{t-1} = i \mid y(t)) \\
&= k_t P(y_t \mid S_t = j, S_{t-1} = i, y(t-1)) P(S_t = j \mid S_{t-1} = i, y(t-1)) P(S_{t-1} = i \mid y(t-1)) \\
&= k_t \sqrt{\frac{1}{2\pi V_e^{(i,j)}}} \exp\left[\frac{-\left(y_t - m_{t-1}^{(i)} - b_{t-1}^{(i)}\right)^2}{2V_e^{(i,j)}}\right] \pi_j q_{t-1}^{(i)},
\end{aligned}$$

where

$$\begin{aligned}
V_e^{(i,j)} &= r_{11}^{(i,j)} + V_\epsilon^{(j)} \\
r_{11}^{(i,j)} &= v_{\mu\mu,t-1}^{(i)} + 2v_{\mu\beta,t-1}^{(i)} + v_{\beta\beta,t-1}^{(i)} + V_\gamma^{(j)} + V_\delta^{(j)}
\end{aligned}$$

and $\pi_j = P(S_t = j \mid S_{t-1} = i, y(t-1))$.

Because the number of normal distributions in the mixture normal distribution of $(\mu_t, \beta_t \mid y_t)$ grows exponentially, they are condensed and the distribution is approximated by a weighted bivariate normal distribution as follows:

$$(\mu_t, \beta_t \mid y_t) \sim \sum_{j=1}^{4} q_t^{(j)} N(\psi_t^{(j)}),$$

where

$$\begin{cases} q_t^{(j)} & = \sum_i p_t^{(i,j)} \\ m_t^{(j)} & = \frac{1}{q_t^{(j)}} \sum_i p_t^{(i,j)} m_t^{(i,j)} \\ b_t^{(j)} & = \frac{1}{q_t^{(j)}} \sum_i p_t^{(i,j)} b_t^{(i,j)} \\ v_{\mu\mu,t}^{(j)} & = \frac{1}{q_t^{(j)}} \sum_i p_t^{(i,j)} \left[v_{\mu\mu,t}^{(i,j)} + \left[m_t^{(i,j)} - m_t^{(j)} \right]^2 \right] \\ v_{\mu\beta,t}^{(j)} & = \frac{1}{q_t^{(j)}} \sum_i p_t^{(i,j)} \left[v_{\mu\beta,t}^{(i,j)} + \left[m_t^{(i,j)} - m_t^{(j)} \right] \left[b_t^{(i,j)} - b_t^{(j)} \right] \right] \\ v_{\beta\beta,t}^{(j)} & = \frac{1}{q_t^{(j)}} \sum_i p_t^{(i,j)} \left[v_{\beta\beta,t}^{(i,j)} + \left[b_t^{(i,j)} - b_t^{(j)} \right]^2 \right] \end{cases}$$

In this way, the posterior at time t is in the same form as the posterior at time $t-1$ so that the same procedure can again be applied at time t and so on.

3.3 Extension of the Harrison & Stevens method

In this section, we extend the Harrison & Stevens method using not only the previous state to infer the next one, but the two previous states. The method can be extended to as many previous states as one wishes and we have done the necessary computation for this extension. But for now, we will focus only on the case of two previous states because an extension to more than two previous states would not only be computationally demanding, but also the results with this restriction are already convenient as we shall see in the next sections.

We know that if we assume at time $t-2$ a mixture of bivariate normal distributions:

$$(\mu_{t-2}, \beta_{t-2} \mid y_{t-2}) \sim \sum_{i=1}^{4} q_{t-2}^{(i)} N(\psi_{t-2}^{(i)}),$$

we can obtain the distribution at time $t-1$ as follows:

$$(\mu_{t-1}, \beta_{t-1} \mid y_{t-1}) \sim \sum_{i,j} p_{t-1}^{(i,j)} N(\psi_{t-1}^{(i,j)}),$$

where the parameters are computed as stated in Section 3.2.

At this point instead of condensing the prior as before, we conduct another step of the update algorithm to infer the distribution at time t by the same algorithm as previously. Assuming for each current state $h = 1, 2, 3$ and 4 that

$$(\mu_t, \beta_t \mid y_t, S_t = h, S_{t-1} = j, S_{t-2} = i) \sim N(\psi_t^{(i,j,h)}),$$

where $\psi_t^{(i,j,h)} = B(\psi_{t-1}^{(i,j)}; V_\epsilon^{(h)}, V_\gamma^{(h)}, V_\delta^{(h)})$, the complete posterior is given by

$$(\mu_t, \beta_t \mid y_t) \sim \sum_{i,j,h} p_t^{(i,j,h)} N(\psi_t^{(i,j,h)}).$$

We just need to compute the probabilities $p_t^{(i,j,h)} = P(S_t = h, S_{t-1} = j, S_{t-2} = i \mid y(t))$. Again let $k_t = 1/P(y_t \mid y(t-1))$ as in the previous section. We can then write

$$\begin{aligned}
p_t^{(i,j,h)} &= P(S_t = h, S_{t-1} = j, S_{t-2} = i \mid y(t)) \\
&= k_t P(y_t \mid S_t = h, S_{t-1} = j, S_{t-2} = i, y(t-1)) \times P(S_t = h \mid S_{t-1} = j, S_{t-2} = i, y(t-1)) \\
&\times P(S_{t-1} = j, S_{t-2} = i \mid y(t-1)) \\
&= k_t \sqrt{\frac{1}{2\pi V_e^{(i,j,h)}}} \exp\left[\frac{-\left(y_t - m_{t-1}^{(i,j)} - b_{t-1}^{(i,j)}\right)^2}{2V_e^{(i,j,h)}}\right] \pi_h p_{t-1}^{(i,j)},
\end{aligned}$$

where by analogy, we define:

$$\begin{aligned}
V_e^{(i,j,h)} &= r_{11}^{(i,j,h)} + V_\epsilon^{(h)} \\
r_{11}^{(i,j,h)} &= v_{\mu\mu,t-1}^{(i,j)} + 2v_{\mu\beta,t-1}^{(i,j)} + v_{\beta\beta,t-1}^{(i,j)} + V_\gamma^{(h)}
\end{aligned}$$

and $\pi_h = P(S_t = h \mid S_{t-1} = j, S_{t-2} = i, y(t-1))$.

At this point, we condense the prior to stop the exponential growth of the number of values to be stored. We apply the same weighting method as previously, which yields the following mixture of bivariate normal distributions:

$$(\mu_t, \beta_t \mid y_t) \sim \sum_{j,h} p_t^{(j,h)} N(\psi_t^{(j,h)}),$$

where

$$\begin{cases} p_t^{(j,h)} &= \sum_i p_t^{(i,j,h)} \\ m_t^{(j,h)} &= \frac{1}{p_t^{(j,h)}} \sum_i p_t^{(i,j,h)} m_t^{(i,j,h)} \\ b_t^{(j,h)} &= \frac{1}{p_t^{(j,h)}} \sum_i p_t^{(i,j,h)} b_t^{(i,j,h)} \\ v_{\mu\mu,t}^{(j,h)} &= \frac{1}{p_t^{(j,h)}} \sum_i p_t^{(i,j,h)} \left[v_{\mu\mu,t}^{(i,j,h)} + \left[m_t^{(i,j,h)} - m_t^{(j,h)} \right]^2 \right] \\ v_{\mu\beta,t}^{(j,h)} &= \frac{1}{p_t^{(j,h)}} \sum_i p_t^{(i,j,h)} \left[v_{\mu\beta,t}^{(i,j,h)} + \left[m_t^{(i,j,h)} - m_t^{(j,h)} \right] \left[b_t^{(i,j,h)} - b_t^{(j,h)} \right] \right] \\ v_{\beta\beta,t}^{(j,h)} &= \frac{1}{p_t^{(j,h)}} \sum_i p_t^{(i,j,h)} \left[v_{\beta\beta,t}^{(i,j,h)} + \left[b_t^{(i,j,h)} - b_t^{(j,h)} \right]^2 \right] \end{cases}$$

In this way, the posterior at time t is in the same form as the posterior at time $t-1$, so that the same procedure can again be applied at time t and so on.
Applying this method allows us not only to infer the state at time t based on all informations from time $t-1$ and $t-2$, but also to compute new probabilities such as

- the probability of being in state i at time $t-2$ given all data until time t,

$$P(S_{t-2} = i \mid y(t)) = \sum_{j,h} p_t^{(i,j,h)} .$$

- the probability of being in state j at time $t-1$ given all data until time t,

$$P(S_{t-1} = j \mid y(t)) = \sum_{i,h} p_t^{(i,j,h)} .$$

3.4 Improvement of the state change detection

3.4.1 Transformation of the data to achieve normality

A common problem in practice is that the observed data are not normally distributed. For example the data could be right skewed like manufacturing data. This is why we opt for a Box-Cox transformation of the data before applying the state change detection procedure, so that our state space model transforms to

$$\begin{aligned}y_t^{(\lambda)} &= \mu_t + \epsilon_t \\ \mu_t &= \mu_{t-1} + \beta_{t-1} + \gamma_t \\ \beta_t &= \beta_{t-1} + \delta_t\,.\end{aligned}$$

The estimation of the transformation parameter λ will be done from a short starting sequence of the data, if historical data is not available. Due to the fact that the starting sequence can contain outliers, it is recommended to use a robust estimate of the transformation parameter. We use the estimator of λ based on our robust asymmetric Shapiro-Wilk test λ_{RSW}.

We investigate the minimal length of the starting sequence for suitable estimation to achieve approximate normality. For this purpose, for each value of λ in $\{0, 0.25, 0.5, 0.75, 1\}$ and for sample sizes ranging from 10 to 300 observations, we simulate 1000 samples from the inverse Box-Cox family, where the mean and the variance of the normal distributions before inverse transformation are 7 and 1, respectively.

Then we perform the robust and the usual Shapiro-Wilk test on the transformed sample with the estimated transformation parameter ($\hat{\lambda}_{RSW}$). The results are reported in Table 3.1. The chosen significance level is 0.05. We also compute the bias and the root mean square error of the estimate of λ and report the results in Table 3.2. We note that the biases of the estimate of λ are not so large for small sample sizes. We remark that the root mean square error is still large in small samples, this can be due to the fact that for its computation, larger estimation errors are more heavily weighted than smaller ones. On the other hand, Table Table 3.1 reports good results for the transformation for small sample sizes, so that we can choose a short starting sequence without affecting the transformation to normality too severely. But of course, the larger the starting sequence the better.

Sample size	Robust Shapiro-Wilk	Shapiro-Wilk
10	98.52	93.08
15	98.90	93.32
20	99.06	94.20
25	98.76	93.86
30	99.02	94.10
35	98.94	93.68
40	98.88	93.96
45	99.20	94.56
50	99.06	94.20
75	99.20	95.04
100	99.24	95.38
125	99.06	95.30
150	99.24	96.08
175	99.06	95.48
200	99.22	96.28
300	99.28	96.46

Table 3.1: Pass rates (in %) of the robust and usual Shapiro-Wilk test applied to the transformed data with $\hat{\lambda}_{RSW}$ at a significance level of 0.05.

3.4.2 Classification of observations to a state

Our major goal is to detect state changes in the data. Although Harrison and Stevens have provided us with a procedure to determine the a posteriori probability of the occurrence of a state, they have not actually classified each observation to a given state. In this section, we provide a classification procedure.

While it is quite obvious that it is difficult to distinguish a level shift from an outlier instantaneously, we can still distinguish a level shift or an outlier from a slope change. In this scope, we investigate two main cases. Firstly, instantaneous classification which

Sample size	Bias	RMSE
10	-0.05	0.49
15	-0.07	0.46
20	-0.07	0.44
25	-0.08	0.43
30	-0.08	0.42
35	-0.08	0.40
40	-0.10	0.41
45	-0.09	0.39
50	-0.09	0.39
75	-0.08	0.35
100	-0.08	0.33
125	-0.07	0.31
150	-0.07	0.29
175	-0.07	0.27
200	-0.06	0.26
300	-0.04	0.22

Table 3.2: Bias and root mean square error of the estimate of the transformation parameter λ for different sample sizes and true value of λ in $\{0, 0.25, 0.5, 0.75, 1\}$.

is done at the same time as the data becomes available, and secondly the one-step-after classification that is done one time point after a data point is observed. This method can be extended to as many time lags before classification as one wishes, but we focus only on the previously mentioned cases.

We opted for the linear discriminant analysis for classification. To determine the classification rules, we perform an extensive simulation and find the results to be quite appealing.

3.4.2.1 Settings for the simulation

To find a classification rule, we conducted intensive simulations. We simulate 100000 time series of length 100 for our classification. Among these, there were 25000 time series with no change, that is the time series consists only of steady states along all the 100 observations in each time series. The remaining 75000 time series are divided into 3 groups representing the 3 remaining states (step change, slope change and outlier). We generate for each of these states 25000 time series and build in each of these 25000 time series at a given time a state change. This means, for example, that 25000 time series were generated for the step change. In each of the 25000 time series we built one step change of different magnitudes at a randomly chosen point in time. The other states were treated in the same manner.

To simulate a time series as close as possible to reality, we chose a set of parameters to suit our goals. Recall the basic model

$$\begin{aligned} y_t &= \mu_t + \epsilon_t \\ \mu_t &= \mu_{t-1} + \beta_{t-1} + \gamma_t \\ \beta_t &= \beta_{t-1} + \delta_t, \end{aligned}$$

where ϵ_t, γ_t and δ_t are independent normally distributed random variables with mean 0 and variance σ_ϵ^2, σ_γ^2 and σ_δ^2, respectively. We simulated each time series by setting the variances σ_ϵ^2, σ_γ^2 and σ_δ^2 to 1, 10^{-4} and 10^{-8}, respectively. We find these settings of the variance to give realistic time series. Because we need starting values μ_0 and β_0, we opted for a value of μ_0 in the interval $[25, 50]$, and β_0 was chosen to be either 0 with a probability of 0.9 or 0.05 with a probability of 0.1. The lower bound of μ_0 is chosen for y_t to be non-negative, and the upper bound so that the process y_t stays within a reasonable range, especially when we will use the Box-Cox parameter to perform a reverse transformation on it. Most of the times $\beta_0 = 0$ suits reality well, but to incorporate some time series with a positive slope at the beginning, we also chose a random positive but not too large value for this parameter as stated above. The simulated process is in a steady state and normally distributed.

To simulate the state changes, we chose randomly a time between 30 and 90 and we created state changes of different magnitudes at that chosen time. We generated step changes with a magnitude chosen randomly in between $500\sigma_\gamma$ and $2000\sigma_\gamma$. The slope changes magnitudes lie between $10000\sigma_\delta$ and $50000\sigma_\delta$. Finally the magnitude of an outlier is chosen randomly between $4\sigma_\epsilon$ and $10\sigma_\epsilon$.

Since the so created process is normally distributed, we made an inverse Box-Cox transformation to obtain a process to which we will apply the state change detection procedure. We set the transformation parameter λ to take one of the four values 0.25, 0.5, 0.75 or 1 randomly.

3.4.2.2 Instantaneous classification

The naive idea would be to classify the observations according to the estimated a posteriori probabilities of occurrence of a state. In the case of an instantaneous classification, this would yield a misclassification rate of 5.35%. We will see that the linear discriminant analysis yields better results.

To deduce an instantaneous classification rule when the data at time t arrives, we used 7 variables:

- The estimated probability of an outlier at time t given all data until time t named bf04
- The observations y_{t-2}, y_{t-1} and y_t
- The estimated value of λ
- The estimated standard deviation of the observation noise ϵ, $\hat{\sigma}_\epsilon$
- The ratio between the difference $y_t - y_{t-1}$ and the estimated standard deviation of the observation noise, $\dfrac{y_t - y_{t-1}}{\hat{\sigma}_\epsilon}$ named ry_{0_b}.

The estimated probability of a steady state will always be left out because the probabilities of the states sum to 1, and with 3 of them fixed the last one is known. It should also be noticed that the estimated probabilities of a step and slope change at time t given all the data until time t were left out. The reason is that we found out that when one uses the ratio of variances suggested by Harrison & Stevens (1971), the a posteriori probabilities of the step and slope changes are always correlated to the a posteriori probability of a transient. These variables were omitted to avoid collinearity between variables while performing linear discriminant analysis. Also note the denomination bf04 for example, that stands for (b)Bayes (f)Factor (0)instantaneous for an (4)outlier. Instead of using the Bayes Factor, the posterior probability was used since it is proportional to the Bayes Factor and it would not change the results of the classification.

We coded the classes so that the steady state and the slope change are in the same class coded 0, while the step change and the outlier are coded 1. We coded the states in this manner, since given the information until time t, it is quite impossible to distinguish between a step change and an outlier. On the other hand, the steady state and the slope change are in the same class due to the fact that it is very unlikely to detect a slope change instantaneously, since if a slope change occurs the next observation will not differ too much from a steady state, unless the magnitude of the slope change is incredibly large, and in this case the observation will be instantaneously classified as outlier or step change. We randomly select 90% of the data as training set and the rest as test sample. A variable selection procedure run with a linear discriminant analysis shows that the model with the smallest training classification error is the model given in Table 3.3, which also contains the estimated coefficients of the linear discriminant. This model yields an out-of-sample classification error of 2.1%. Note that the Intercept in Table 3.3 is computed so that the training error is minimized as suggested by Hastie et al. (2009, page. 111).

Variables	Coefficients
bf 04	8.45
y_{t-1}	-3.29×10^{-6}
ry_{0_b}	9.37×10^{-7}
Intercept	-2.914

Table 3.3: Coefficients of the linear discriminants for the model for the instantaneous classification.

A new vector $x^* = (bf04^*, y^*_{t-1}, ry^*_{0\_b})$ is classified into class 1 (outlier or step change) if the following rule applies

$$8.45\, bf04^* - 3.29 \times 10^{-6}\, y^*_{t-1} + 9.37 \times 10^{-7}\, ry^*_{0\_b} - 2.914 \geq 0$$

and to class 0 (steady state or slope change) otherwise.

3.4.2.3 One-step-after classification

In the previous paragraph, we discussed the instantaneous classification of an observation making no distinction between the step change and the transient. It makes sense after one further observation to try to dissociate these two states.

If we opt for the naive classification, which classifies an observation to the class with largest one-step-after a posteriori probability, the misclassification error would be 7.29%. To improve the classification, we derive a classification rule applying the linear discriminant analysis to our simulated data.

For this purpose, 4 variables were added to the 7 previous ones:

- The probability of a step change at time t given all the data until time $t+1$ denoted by bf12.
- The probability of a slope change at time t given all the data until time $t+1$ denoted by bf13.

- The probability of an outlier at time t given all the data until time $t+1$ denoted by bf14.

- The ratio of the difference $y_{t+1} - y_{t-1}$ and the estimated standard deviation of the observation noise, $\dfrac{y_{t+1} - y_{t-1}}{\hat{\sigma}_\epsilon}$ denoted by ry_{a_b}.

Now we have 11 variables to perform the linear discriminant analysis. For the purpose of this classification, we code the outliers as 2 and the step changes as 1, while the slope change and the steady states were coded as 0. In the same manner, we used variable selection and misclassification rate to select the best model for our data. We compute the intercept in the same manner as for the instantaneous classification. The best model is given in Table 3.4. The linear discriminant analysis with these variables yields an out-of-sample classification error for discrimination between steady state, slope change, level shift and transient of 3.11%.

Variables	Coefficients of linear discriminant 1	Coefficients of linear discriminant 2
bf14	-2.87	9.37
bf04	10.59	-5.51
y_{t-1}	-2.55×10^{-5}	5.91×10^{-5}
y_t	3.48×10^{-6}	3.02×10^{-5}
y_{t+1}	2.06×10^{-5}	-8.19×10^{-5}
$\hat{\sigma}_\epsilon$	-2.39×10^{-4}	-2.02×10^{-4}
ry_{a_b}	-1.21×10^{-6}	2.40×10^{-6}
Intercept	-3.65	-3.57

Table 3.4: Coefficients of the linear discriminants for the model for the one-step-after classification.

Since there are 3 classes this time, we can state the classification rule in 2 steps. Given a new observation $x^* = (bf14^*, bf04^*, y^*_{t-1}, y^*_t, y^*_{t+1}, \hat{\sigma}^*_\epsilon, ry^*_{a\_b})^t$, we have:

Step one

The second linear discriminant separates the classes 0 (steady state and slope change) and 1 (step change) from the class 2 (outlier). If we have

$$\begin{aligned} &9.37\, bf14^* - 5.51\, bf04^* + 5.91 \times 10^{-5}\, y^*_{t-1} + 3.02 \times 10^{-5}\, y^*_t - 8.19 \times 10^{-5}\, y^*_{t+1} \\ &- 2.02 \times 10^{-4}\, \hat{\sigma}^*_\epsilon + 2.4 \times 10^{-6}\, ry^*_{a\_b} - 3.57 \geq 0, \end{aligned} \tag{3.5}$$

then classify the observation to class 2, otherwise the class will be either 1 or 0. This leads us to the second step.

Step two

The first linear discriminant component distinguishes the class 0 (steady state and slope change) from 1 (step change). If the following holds

$$\begin{aligned} &- 2.87\, bf14^* + 10.59\, bf04^* - 2.55 \times 10^{-5}\, y^*_{t-1} + 3.48 \times 10^{-6}\, y^*_t + 2.06 \times 10^{-5}\, y^*_{t+1} \\ &- 2.39 \times 10^{-4}\, \hat{\sigma}^*_\epsilon - 1.21 \times 10^{-6}\, ry^*_{a\_b} - 3.65 \geq 0, \end{aligned} \tag{3.6}$$

then the observation is assigned to class 1, that is step change, and otherwise to class 0.

3.4.2.4 Classification of the slope change

Let $y_1, \ldots, y_n$ be an observed stretch of a time series (y_t). We recall the state space model with a local linear trend (3.1)

$$\begin{aligned} y_t &= \mu_t + \epsilon_t \\ \mu_t &= \mu_{t-1} + \beta_{t-1} + \gamma_t \\ \beta_t &= \beta_{t-1} + \delta_t\,. \end{aligned}$$

Here ϵ_t, γ_t and δ_t are independent normally distributed random variables with mean 0 and variances σ^2_ϵ, σ^2_γ and σ^2_δ, respectively.

Let us define β_0 and μ_0 as initial values for the processes β_t and μ_t. Then we can write:

$$\beta_t = \beta_{t-1} + \delta_t \Rightarrow \beta_t = \beta_0 + \sum_{i=1}^{t} \delta_i, \quad \text{for} \quad t \geqslant 1 \tag{3.7}$$

$$\begin{aligned}
\mu_t = \mu_{t-1} + \beta_{t-1} + \gamma_t \Rightarrow \mu_t &= \mu_0 + \sum_{i=0}^{t-1} \beta_i + \sum_{i=1}^{t} \gamma_i \\
\Leftrightarrow \mu_t &= \mu_0 + \sum_{i=1}^{t-1} \left(\beta_0 + \sum_{j=1}^{i} \delta_j \right) + \beta_0 + \sum_{i=1}^{t} \gamma_i \\
\Leftrightarrow \mu_t &= \mu_0 + \sum_{i=0}^{t-1} \beta_0 + \sum_{i=1}^{t-1} \sum_{j=1}^{i} \delta_j + \sum_{i=1}^{t} \gamma_i \\
\Leftrightarrow \mu_t &= \mu_0 + t\beta_0 + \sum_{i=1}^{t-1} (t-i)\delta_i + \sum_{i=1}^{t} \gamma_i, \quad \text{for} \quad t \geqslant 1\,.
\end{aligned} \tag{3.8}$$

This yields the following equation for the observed process y_t:

$$y_t = \mu_t + \epsilon_t \Rightarrow y_t = \mu_0 + t\beta_0 + \sum_{i=1}^{t-1} (t-i)\delta_i + \sum_{i=1}^{t} \gamma_i + \epsilon_t, \quad \text{for} \quad t \geqslant 1 \tag{3.9}$$

If we assume that at a given time τ a slope change of magnitude ω occurs, we can replace the noise δ_t by

$$\Delta_t = \begin{cases} \delta_t, & \text{if } t \neq \tau \\ \delta_\tau + \omega, & \text{if } t = \tau\,. \end{cases}$$

Therefore writing the new model as follows

$$\begin{aligned}
y_t &= \mu_t + \epsilon_t \\
\mu_t &= \mu_{t-1} + \beta_{t-1} + \gamma_t \\
\beta_t &= \beta_{t-1} + \Delta_t\,,
\end{aligned}$$

we can see that equations (3.7), (3.8) and (3.9) hold until time $\tau - 1$. After time $t = \tau$, we have

$$\beta_t = \beta_0 + \sum_{i=1}^{t} \Delta_i = \omega + \beta_0 + \sum_{i=1}^{t} \delta_i, \quad \text{for} \quad t \geqslant \tau\,,$$

so that in the process (β_t) there is a step change of magnitude ω. Incorporating this into the expression of μ_t will only affect the process after time $t = \tau + 1$ giving the new expression

$$\mu_t = (t-\tau)\omega + \mu_0 + t\beta_0 + \sum_{i=1}^{t-1}(t-i)\delta_i + \sum_{i=1}^{t}\gamma_i, \quad \text{for} \quad t \geqslant \tau + 1\,.$$

This expression shows a linear increase of magnitude ω after each observation starting at time $t = \tau + 1$ in the observation mean μ_t.

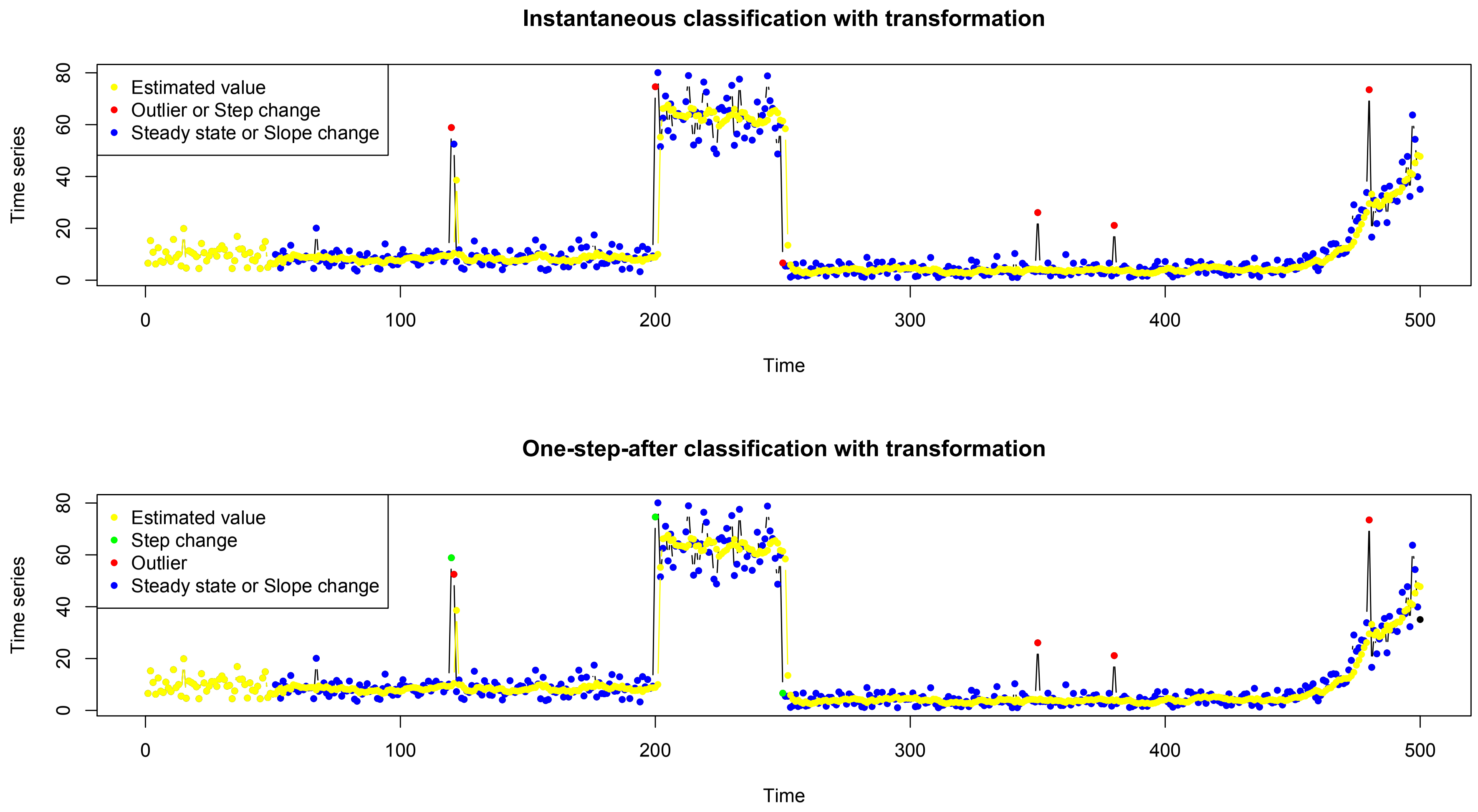
A natural consequence is that the detection of a slope change depends on the relationship between the magnitude of the slope change ω and the standard deviation of the noise of the observation σ_ϵ. Since the observation noise is assumed to be normally distributed, we can say that until the product $(t-\tau)\omega$ falls out of a reasonable range, the slope will be difficult to detect or the detection will be delayed until a sufficiently large number of future observations is affected, depending on the values of the noises around the time of occurrence of the slope change.

3.5 Example

To show how the extended and improved state change detection procedure works, we simulate a time series of length 500 from state space model (3.1) and build

- 5 outliers at times 120, 121, 350, 380 and 480 of magnitude 9, 8, 6, 5 and 7 standard deviations of ϵ_t respectively.
- 2 step changes at times 200 and 250 of magnitude 1000 and -1200 standard deviations of γ_t, respectively, and corresponding to 10 and -12 standard deviations of ϵ_t respectively.
- One slope change at time 450 of magnitude 2000 standard deviations of δ_t corresponding to 0.2 standard deviation of ϵ_t.

Then we transform the time series with the inverse Box-Cox transformation with true transformation parameter $\lambda = 0.5$ and perform the state change detection with and without Box-Cox transformation.



Instantaneous classification with transformation

One-step-after classification with transformation

Figure 3.2: Classification of the state change detection procedure with robust transformation to normality.

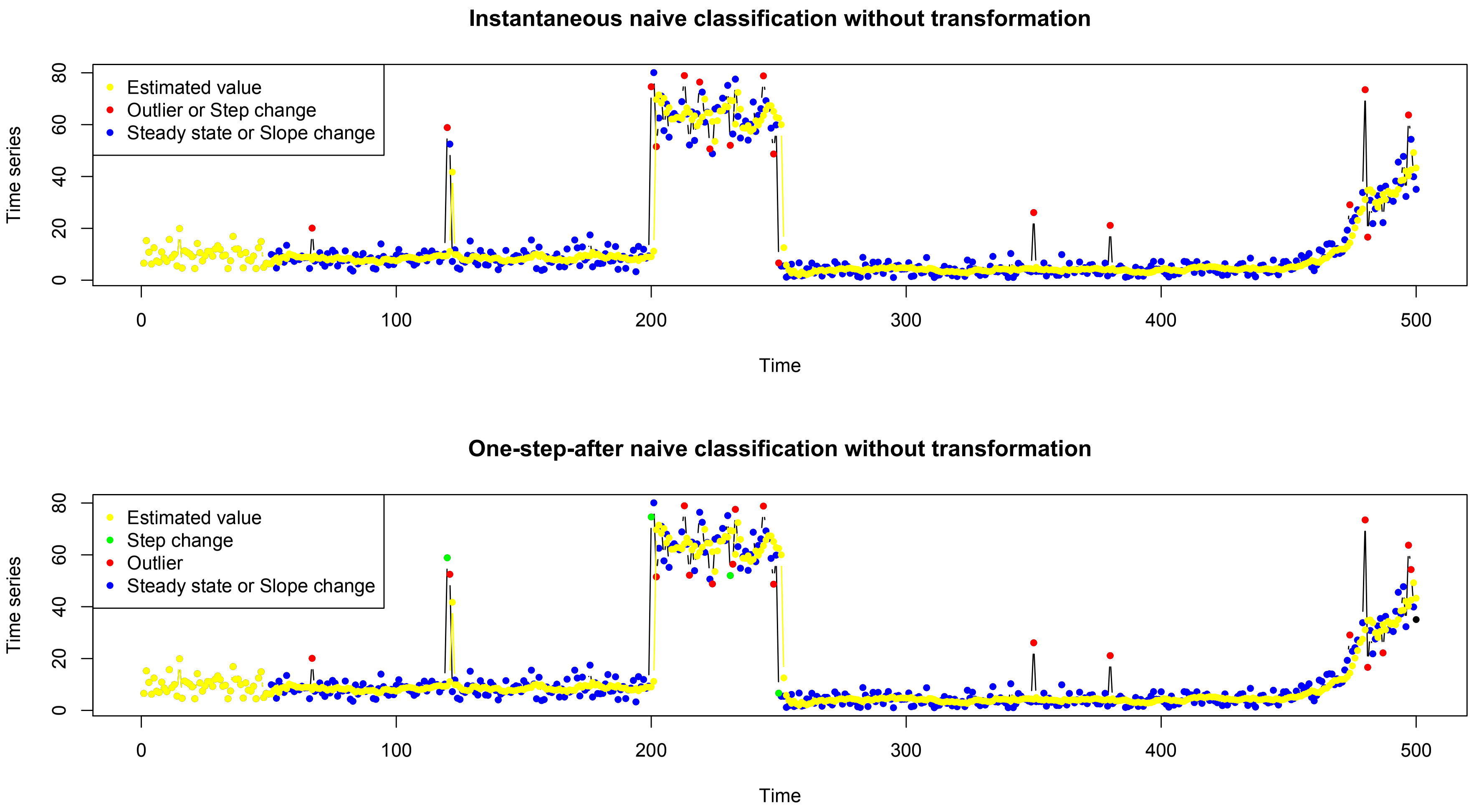


Figure 3.3: Classification of the state change detection procedure without robust transformation to normality.

The robust transformation parameter is estimated from a start sequence of length 50. Figure 3.2 illustrates the classification using the previously determined rules with transformation. While in Figure 3.3, the naive classification is reported, when the Box-Cox transformation i not performed. We remark that the robust transformation enhances the performance of the state change classification procedure, since the classification without transformation puts a lot of steady state observations in the class of outliers instantly and one step after. Tables 3.5 and 3.6 contain the estimated instantaneous and one-step-after a posteriori probabilities, respectively, of each state, computed after robust transformation and without robust transformation. We have just presented the values around the incorporated state changes. We can see that the results with and without transformation are quite similar around the state changes. But as we can see for time 481, where we did not create a state change, the procedure without robust transformation misclassifies this observation as outlier as we have already seen in Figure 3.3 for several other points. In other words, the robust transformation is a very useful and necessary improvement of the state change detection procedure.

3.6 Conclusion

The primary objective of this chapter is to detect state changes in time series. We consider four states : steady state (normal state), step change (level shift), slope change and outlier. For this purpose, we chose the state change detection procedure of Harrison & Stevens (1971), which is a Bayesian method using the Kalman Filter to fit a state space model with a local linear trend. At time $t-1$, for each state change, an a priori distribution for the parameter is specified and the a posteriori probabilities of occurrence of each state at time t are computed. Firstly, we rectified the update equations used in the procedure. Secondly, we extended the procedure to compute the probability of a state change at time $t-2$ given all data until time t and used this feature later on in our improvement process. Because the procedure is based on the assumption of normality, we used our robust Box-Cox transformation to transform the data to approximate normality, hereby widening the scope of application of the procedure to skewed data, for example. A major improvement is the classification of an observation to a state, because the procedure only estimates the probability of occurrence of a state at a given time based on all the information available up to that time, but no classification is specified. By using linear discriminant analysis, we derive an instantaneous classification, which differentiates the steady state and slope change from the step change and the outlier at the arrival of each observation. The one-step-after classification then distinguishes the step change from the outlier one step after the arrival of an observation. An example shows that the transformation enhances the classification procedure and that without transformation a lot of ordinary observations are misclassified as outliers. It is also shown that our method outperforms the naive classification to the class with the largest computed a posteriori probabilities of occurrence.

		Estimated a posteriori probabilities with transformation				Estimated a posteriori probabilities without transformation			
true state	time	steady	step	slope	outlier	steady	step	slope	outlier
steady	119	0.99	0.00	0.00	0.01	0.99	0.00	0.00	0.01
outlier	**120**	0.00	0.03	0.00	0.97	0.00	0.03	0.00	0.97
outlier	**121**	0.70	0.01	0.00	0.29	0.67	0.01	0.00	0.32
steady	122	0.96	0.00	0.00	0.04	0.97	0.00	0.00	0.03
steady	199	0.99	0.00	0.00	0.01	0.99	0.00	0.00	0.01
step	**200**	0.00	0.03	0.00	0.97	0.00	0.03	0.00	0.97
steady	201	0.75	0.01	0.00	0.24	0.86	0.00	0.00	0.13
steady	249	0.99	0.00	0.00	0.01	0.96	0.00	0.00	0.03
step	**250**	0.00	0.03	0.00	0.97	0.00	0.03	0.00	0.97
steady	251	0.74	0.01	0.00	0.25	0.86	0.00	0.00	0.13
steady	349	0.98	0.00	0.00	0.01	0.99	0.00	0.00	0.01
outlier	**350**	0.00	0.03	0.00	0.97	0.00	0.03	0.00	0.97
steady	351	0.98	0.00	0.00	0.01	0.99	0.00	0.00	0.01
steady	379	0.97	0.00	0.00	0.03	0.98	0.00	0.00	0.01
outlier	**380**	0.01	0.03	0.00	0.96	0.00	0.03	0.00	0.96
steady	381	0.98	0.00	0.00	0.01	0.98	0.00	0.00	0.01
steady	449	0.98	0.00	0.00	0.02	0.98	0.00	0.00	0.01
slope	**450**	0.98	0.00	0.00	0.01	0.99	0.00	0.00	0.01
steady	451	0.98	0.00	0.00	0.01	0.99	0.00	0.00	0.01
steady	479	0.98	0.00	0.00	0.02	0.97	0.00	0.00	0.02
outlier	**480**	0.02	0.03	0.00	0.95	0.00	0.03	0.00	0.97
steady	481	0.76	0.01	0.01	0.22	0.03	0.03	0.00	0.94

Table 3.5: Instantaneous estimated a posteriori probabilities of state changes with and without transformation to normality.

		Estimated a posteriori probabilities with transformation				Estimated a posteriori probabilities without transformation			
true state	time	steady	step	slope	outlier	steady	step	slope	outlier
steady	119	0.99	0.00	0.00	0.01	0.99	0.00	0.00	0.01
outlier	**120**	0.00	0.71	0.00	0.29	0.00	0.69	0.00	0.31
outlier	**121**	0.03	0.00	0.00	0.97	0.01	0.00	0.00	0.99
steady	122	0.99	0.00	0.00	0.01	0.99	0.00	0.00	0.01
steady	199	0.99	0.00	0.00	0.01	0.99	0.00	0.00	0.01
step	**200**	0.00	0.77	0.00	0.23	0.00	0.89	0.00	0.11
steady	201	0.93	0.01	0.00	0.06	0.89	0.00	0.00	0.10
steady	249	0.99	0.00	0.00	0.01	0.96	0.00	0.00	0.03
step	**250**	0.00	0.76	0.00	0.24	0.00	0.88	0.00	0.12
steady	251	0.97	0.01	0.00	0.02	0.98	0.00	0.00	0.01
steady	349	0.98	0.00	0.00	0.01	0.99	0.00	0.00	0.01
outlier	**350**	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
steady	351	0.99	0.00	0.00	0.01	0.99	0.00	0.00	0.01
steady	379	0.97	0.00	0.00	0.03	0.98	0.00	0.00	0.01
outlier	**380**	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
steady	381	0.99	0.00	0.00	0.01	0.99	0.00	0.00	0.01
steady	449	0.98	0.00	0.00	0.02	0.98	0.00	0.00	0.01
slope	**450**	0.99	0.00	0.00	0.01	0.99	0.00	0.00	0.01
steady	451	0.99	0.00	0.00	0.01	0.99	0.00	0.00	0.01
steady	479	0.97	0.00	0.01	0.02	0.98	0.00	0.00	0.02
outlier	**480**	0.01	0.01	0.00	0.99	0.00	0.01	0.00	0.99
steady	481	0.81	0.00	0.00	0.18	0.05	0.00	0.00	0.95

Table 3.6: One-step-after probabilities of state changes with and without transformation to normality.

Summary and outlook

The primary objective of this thesis is the construction of a powerful state change detection procedure for monitoring time series, which can help decision makers to react faster to changes in the system and define the proper course of action for each case.

Without losing sight of our primary goal, we first derived a robust test of approximate normality based on the Shapiro-Wilk test (RSW), which detects if the majority of the data follows a normal distribution. The RSW test is based on the idea of trimming the original sample, and replacing the observations in the tail by artificially generated normally distributed data, and then performing the Shapiro-Wilk test on the modified sequence. We show that under the null hypothesis of normality the modified sequence is asymptotically normally distributed and that the RSW test statistic has the same asymptotic null distribution as the Shapiro-Wilk test statistic. The RSW test proves to be resistant to outliers and outperforms the other considered robust test for normality in the presence of outliers. Intending to use the RSW test to create a robust estimator of the Box-Cox transformation, we also investigate its behaviour with respect to the inverse Box-Cox transformation. It proves to be resistant to outliers in this case and also outperforms its competitors in presence of a few outliers.

Secondly, we used the RSW test to derive a robust estimator of the Box-Cox transformation parameter ($\hat{\lambda}_{RSW}$). This conforms to the fact that the Box-Cox transformation only achieves approximate normality and the Shapiro-Wilk test of normality is one of the most powerful tests of normality. Gaudard & Karson (2000) already derived a non robust estimator of the Box-Cox transformation parameter based on the Shapiro-Wilk test statistic that outperformed the other estimators considered in their comparison. As expected, $\hat{\lambda}_{RSW}$ is preferable to the maximum-likelihood and the M-estimators (we considered), mainly because it yields a better transformation in the sense that not only are the transformed samples more symmetrical according to the medcouple (a robust measure of symmetry and tail weight), but they also have a higher pass rate for the RSW test and the $MC1$ test at a significance level of 5%.

Finally, returning to the state change detection, we opt for the method of Harrison & Stevens (1971), which considers four states: the steady state (normal state), the step change (level shift), the slope change and the outlier. The assumption of normally distributed data restricts the usage of the procedure, so we transform the data with $\hat{\lambda}_{RSW}$ to achieve approximate normality. We extend the update equations to two observations in the past, that is to compute the probability of occurrence of a state change at time $t-2$ given all available data until time t. This extension is used when we derive classifications rules for the incoming observations, given that the procedure only computes a posteriori probabilities for the different states and does not classify them. We use linear discriminant analysis and intensive simulations to derive the classification rules. We derived an instantaneous classification separating the step change and the outlier from the slope change and the steady state at the arrival of each observations and a one-step-after classification that separates the three classes outlier, step change and slope change, steady state one step after each observation is available. The simulations show that the first rule has an out-of-sample classification error of 2.1% and the second rule 3.11%. Opposed to this, the naive classification rule, which is to classify according to the estimated a posteriori probability, yields misclassification errors of 5.35% and 7.29%, respectively.

Unfortunately, a classification rule for the slope change is not derived. One could take advantage of the fact that information on the past can be extended to as many observations in the past one wishes, increasing the probability of detecting a slope change. In addition, we do not consider other classification procedures than the linear discriminant analysis, although it is possible for other classification procedures to yield better results than ours.

For all the computations in this work, we used the software package R Core Team (2012).

Bibliography

Alvarez-Esteban, P. C., Del Barrio, E., Cuesta-Albertos, J. A., & Matrán, C. (2010). Assessing when a sample is mostly normal. *Computational Statistics & Data Analysis*, *54*(12), 2914–2925.

Andrews, D. F. (1971). A note on the selection of data transformations. *Biometrika*, *58*(2), 249–254.

Atkinson, A. C. (1973). Testing transformations to normality. *Journal of the Royal Statistical Society. Series B (Methodological)*, *35*(3), 473–479.

Bickel, P. J., & Doksum, K. A. (1981). An analysis of transformation revisited. *Journal of the American Statistical Association*, *76*(374), 296–311.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*(2), 211–252.

Brys, G., Hubert, M., & Struyf, A. (2004a). A robust measure of skewness. *Journal of Computational and Graphical Statistics*, *13*(4), 996–1017.

Brys, G., Hubert, M., & Struyf, A. (2004b). A robustification of the Jarque-Bera test of normality. *Symposium A Quarterly Journal In Modern Foreign Literatures*, (pp. 753–760).

Carroll, R. J. (1980). A robust method for testing transformations to achieve approximative normality. *Journal of the Royal Statistical Society. Series B (Methodological)*, *42*(1), 71–78.

Carroll, R. J. (1982). Two examples of transformations when there are possible outliers. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *31*(2), 149–152.

Cesàro, E. (1888). Sur la convergence des séries. *Nouvelles annales de mathématiques*, *Series 3*(7), 49–59.

Chernoff, H. (1952). A measure of asymptotic efficiency of tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, *23*, 493–507.

Chung, S., Pearn, W., & Yang, Y. (2007). A comparison of two methods for transforming non-normal manufacturing data. *The international Journal of Advanced manufacturing Technology*, *31*(9-10), 957–968.

David, H. A., & Nagaraja, H. N. (2003). *Order Statistics*. Wiley series in probability and statistics. Wiley-Interscience, third ed.

Donoho, D. L. (1982). Breakdown properties of multivariate location estimators. Ph.D. Qualifying paper, Harvard University, Boston.

Frishman, F. (1975). On the Arithmetic Means and Variances of Products and Ratios of Random Variables. In Patil, G.P. and Kotz, S. and Ord, J.K. (Ed.) *A Modern Course on Statistical Distributions in Scientific Work*, vol. 17 of *NATO Advanced Study Institutes Series*, (pp. 401–406). Springer Netherlands.

Gaudard, M., & Karson, M. (2000). On estimating the box-cox transformation to normality. *Communications in Statistics - Simulation and Computation*, *29*(2), 559–582.

Gel, Y., & Gastwirth, J. (2008). A robust modification of the Jarque-Bera test of normality. *Economics Letters*, *99*(1), 30–32.

Gel, Y., Miao, W., & Gastwirth, J. (2007). Robust directed tests of normality against heavy-tailed alternatives. *Computational Statistics & Data Analysis*, *51*(5), 2734–2746.

Harrison, P. J., & Stevens, C. F. (1971). A bayesian approach to Short-Term Forecasting. *Operational Research Quarterly*, *22*(4), 341–362.

Harvey, A. C. (1991). *Forecasting structural time series models and the Kalman filter*. Cambridge University Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, second ed.

Hinkley, D. V. (1975). On power transformations to symmetry. *Biometrika*, *62*(1), 101–111.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1985). *Exploring data tables, trends and shapes*. New York: John Wiley and Sons.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, *58*(301), 13–30.

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

Hubert, M., & Van der Veeken, S. (2008). Outlier detection for skewed data. *Journal of chemometrics*, *22*(3-4), 235–246.

Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational Statistics & Data Analysis*, *52*(12), 5186 – 5201.

Jarque, C. M., & Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, *6*(3), 255 – 259.
URL `http://www.sciencedirect.com/science/article/pii/0165176580900245`

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous Univariate Distributions*, vol. 2. Wiley-Interscience, second ed.

Leslie, J. R., Stephens, M. A., & Fotopoulos, S. (1986). Asymtotic distribution of the Shapiro-Wilk W for testing for normality. *The Annals of Statistics*, *14*(4), 1497–1506.

R Core Team (2012). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. URL `http://www.R-project.org/`

Royston, P. (1995). Remark AS R94: A Remark on Algorithm AS 181: The W-test for Normality. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *44*(4), 547–551.

Saskia, R. M. (1992). The box-cox transformation technique: A review. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *41*(2), 169–178.

Serfling, R., & Mazumder, S. (2009). Exponential probability inequality and convergence results for the median absolute deviation and its modifications. *Statistics & Probability Letters*, *79*(16), 1767–1773.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. New York: Wiley.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality. *Biometrika*, *52*(3-4), 591–611.

Stahel, W. A. (1981). *Robuste Schätzungen: infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen*. Ph.D. thesis, ETH Zürich.

Stephens, M. A. (1975). Asymptotic properties for covariance matrices of order statistics. *Biometrika*, *62*(1), 23–28.

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Dissertation selbständig verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Dissertation ist bisher keiner anderen Fakultät vorgelegt worden. Ich erkläre, dass ich bisher kein Promotionsverfahren beendet habe und dass keine Aberkennung eines bereits erworbenen Doktorgrades vorliegt.

Dortmund, 12. April 2013

Arsene Ntiwa Foudjo