Robust change-point detection and dependence modeling

Dissertation

zur Erlangung des Grades „Doktor der Naturwissenschaften“ an der Fakultät Statistik der
Technischen Universität Dortmund
vorgelegt von

Alexander Dürre

Betreuer und Gutachter	Prof. Dr. R. Fried
Betreuer	Dr. D. Vogel
Gutachter	Prof. Dr. Ch. H. Müller
Kommissionsvorsitz	Prof. Dr. M. Wagner
Abgabe der Dissertation:	27.4.2017
Tag der mündlichen Prüfung:	19.7.2017

Contents

1 Introduction 4

2 Robust estimation of the autocorrelation function 7

2.1 Background and Notation 7

2.2 Robust autocorrelation estimators 10

2.2.1 Estimation based on univariate transformations 10

2.2.2 Estimation based on signs and ranks 11

2.2.3 Estimation based on partial autocorrelation 12

2.2.4 Estimation based on multivariate correlation 13

2.2.5 Estimation based on variances 15

2.2.6 Estimation based on robust filtering 16

2.2.7 Positive-semidefiniteness 17

2.3 Simulations 18

2.3.1 Efficiency for uncontaminated data 20

2.3.2 Robustness under contamination 20

2.3.3 Non-linear models 23

2.3.4 Positive-semidefiniteness 28

2.4 Conclusion 28

3 Spatial sign correlation 29

3.1 Introduction 29

3.2 The spatial sign covariance matrix 30

3.3 A spatial sign based estimator for the correlation coefficient 34

3.4 The two-stage spatial sign correlation 37

3.4.1 Asymptotic results 39

3.4.2 Simulations 42

3.5 Comparison of robust correlation estimators 44

3.5.1 Nonparametric estimators 46

3.5.2 GK estimators 47

3.5.3 Affine equivariant estimators 47

3.5.4 Simulations 48

3.6 Multivariate spatial sign correlation matrix 52

3.7 Conclusion 56

3.8 Proofs 57

4 Robust change-point detection in panel data 69

4.1 Introduction 69

4.2 Testing procedure 70

4.3 Theoretical results 73

4.4 Simulation 79

4.5 Summary 81
4.6 Proofs 82

Chapter 1

Introduction

This doctoral thesis is cumulative, consisting of three parts: robust estimation of the autocorrelation function, the spatial sign correlation, and robust change point detection in panel data. Albeit covering quite different statistical branches like time series analysis, multivariate analysis and change point detection, there is a common issue in all the three sections and this is robustness. Robustness is in the sense that the statistical analysis should stay reliable if there is a small fraction of observations which do not follow the chosen model.

The term "robustness" was coined by Box (1953), who originally used it to express insensitivity to non-normality, but the roots of robust statistics are much older. Already Bernoulli (1777) discusses rules to reject outliers and mentions that it is common practice for astronomers to ignore "*observations which they judge to be too wide of the truth.*" Even earlier in 1757 Boscovich estimates the ellipticity of the earth by a trimmed mean due to implausible values (see for example chapter 1 in Koenker, 2005). An extensive and diverting overview about the early stage of robust estimation can be found in Stigler (1973). In the 1960s, several very significant articles appeared, which continue to have strong influence still on today's research. Therefore one can maybe call this decade the starting point of modern robust statistics. It started with the contribution of Tukey (1960), who advocates to focus more on "*robustness of efficiency*". This was a reaction to a number of articles investigating and eventually confirming the validity of some Gaussian procedures under non-normality (see, e.g., Pearson, 1931; Eden and Yates, 1933; Hey, 1938), which is often caused by the central limit theorem. Tukey on the other hand using the example of variance estimation showed that the efficiency of the empirical variance decreases drastically even under "mild" deviations from normality. Four years later Huber (1964) introduced the concept of M-estimation. Originally proposed for the location problem the underlying concept spread over all kinds of statistical applications and is also a key ingredient in this thesis. Two of the most fundamental measurements of robustness were also derived, the influence function in the PhD thesis of Hampel (1968) and the breakdown point in Hodges Jr (1967) respectively Hampel (1971).
In the following robust methods for more complex data structures were developed, like regression problems (Huber, 1973; Mallows, 1975; Hampel, 1975; Siegel, 1982), multivariate data (Maronna and Yohai, 1976; Stahel, 1981; Donoho, 1982; Rousseeuw, 1985), time series (Fox, 1972; Masreliez, 1975; Denby and Martin, 1979; Martin and Yohai, 1986) and infinite dimensional data (Locantore et al., 1999; Fraiman and Muniz, 2001; Bramati and Croux, 2007; Bali et al., 2011), to name only a few. This thesis covers problems of the last three topics.

The *first part* of this thesis is a review study comparing different proposals for robust estimation of the autocorrelation function. Over the years many estimators have been proposed but thorough comparisons are missing, resulting in a lack of knowledge which estimator is preferable

in which situation. We treat this problem, though we mainly concentrate on a special but nonetheless very popular case where the bulk of observations is generated from a linear Gaussian process. This part of the thesis nearly coincides with Dürre, Fried and Liboschik (2015a).
The origin of this work goes back to an IASC summer school in Leuven 2011 where I was yet studying for my masters degree. Motivated by the the time series and robust statistics courses in Dortmund I was interested in the question: "How can one robustly estimate the fundamental autocorrelation function?" When I asked my later doctoral advisor Roland Fried for whom I was working as student assistant at this time, he smiled since he asked himself the self question some years ago. He even had started some very elementary research on this question and gave me notes about it. At the same time my friend Tobias Liboschik, already a PhD student at this time, and I wanted to take part at the summer school and were asked to present something we were working on. Since Tobias' work was barely robust at this time, we decided to talk about robust estimation of the autocorrelation function, searching together for literature, implementing procedures and doing some rudimental simulations just finishing in time for the workshop. The talk was received very well but none of us had time to write an article about it. This changed when I become a PhD student. We originally believed it would require an overseeable effort, but in the end it took us nearly two years to collect more estimators for our comparison and multiple times improve and change our simulations. Tobias deserves special credit for implementations of some estimators and Roland earns recognition for writing the introductory subsection about background and notation. Both of course gave many valuable suggestions and improvement suggestions. Together with the article we also provide an R-package sscor (Dürre, Fried, Liboschik and Rathjens, 2016) containing all considered methods and meanwhile also robust procedures for AR fits, spectrum estimation and change point estimation. Jonathan Rathjens deserves a special mention for the implementation of spectrum functions and Tobias made such a great effort to get the functions easy to use. Up to date the package is only available at R-forge, but it will be available on CRAN soon.

The *second chapter* deals with something congeneric, namely measuring dependence through the spatial sign correlation, a robust and within the elliptic model distribution-free estimator for the correlation based on the spatial sign covariance matrix. We derive its asymptotic distribution and robustness properties like influence function and gross error sensitivity. Furthermore we propose a two stage version which improves both efficiency under normality and robustness. The surprisingly simple formula of its asymptotic variance is used to construct a variance stabilising transformation, which enables us to calculate very accurate confidence intervals, which are also distribution-free within the elliptic model. We also propose a positive semi-definite multivariate spatial sign correlation, which is more efficient but less robust than its bivariate counterpart. Theoretical results regarding the properties of this matrix-valued estimator are challenging and yet largely open problems.
This chapter is based on the articles Dürre, Vogel and Fried (2015b), Dürre and Vogel (2016a), Dürre, Tyler and Vogel (2016) and Dürre, Fried and Vogel (2017). Basically there are two types of changes compared to the referred papers. First I have removed results which arose in collaboration with David E. Tyler and Daniel Vogel which I have not proved myself. Likewise I have excluded parts which derive from research during my bachelor- and master studies. Second, I have streamlined the order of presentation when merging the publications. Not seldom open questions in one article were answered in the following papers.
Also the origins here date back a long time. After I raved about a magnificent analysis lecture where we determined the dimension of the unit sphere with the largest volume, Daniel Vogel, also a PhD student of Roland at this time, gave me the task to solve an integral, which describes the covariance of the bivariate spatial sign covariance matrix. I was able to calculate it by the residue theorem from complex analysis, my favourite mathematical branch. The

result is part of my bachelor thesis „Über die theoretischen Eigenschaften der Spatial-Sign-Kovarianzmatrix im elliptischen Modell.“ After writing my master-thesis („Konsistenzaussagen zur Spatial-Sign-Kovarianzmatrix “) which we published subsequently (Dürre, Vogel and Tyler, 2014) and becoming a PhD student we continued to study the spatial sign covariance matrix. Daniel proposed to devise a correlation estimator, the spatial sign correlation, and found an explicit formula of the estimator based on the empirical spatial sign covariance. I was able to calculate a closed form expression of its asymptotic variance by the inverse function theorem and derived influence function and gross error sensitivity. Results were published in Dürre, Vogel and Fried (2015b).

Since the estimator tends to be inefficient under heteroscedasticity we proposed a two step procedure where the data is first standardized marginally by a robust scale estimator. Both referees of Dürre, Vogel and Fried (2015b) suggested calculation of the influence function of the two step procedure. We complied with this request and it inspired us to a subsequent article about the theoretical properties of the two step spatial sign correlation. With Daniels guidance I was able to prove the asymptotic negligibility of the scale estimation of the pre standardization. Motivated by the result that the variance of the two step procedure only depends on the correlation itself, I remembered Daniels idea to use the spatial sign correlation for testing, which was practically not possible for the one step version, and derived following Fishers z-transform a variance stabilising transformation for the estimator. Daniel did the main part of writing the actual article Dürre and Vogel (2016a).

Daniel then proposed to investigate the symmetrized spatial sign covariance matrix. While looking for some helpful formulas to derive its asymptotic variance in Gradshteyn and Ryzhik (2000) I accidentally stumbled upon a formula which permits to describe the eigenvalues of the spatial sign covariance matrix for arbitrary dimension as one dimensional integrals. These, can be determined fast and accurately by a Gauss-Jacobi quadrature, which is by the way implemented in the R-package Dürre and Vogel (2016b). We published the result together with two Propositions of David E. Tyler in the article Dürre, Tyler and Vogel (2016).

To derive a multivariate spatial sign correlation it remains to invert the relationship between the eigenvalues. I first experimented with a Newton type algorithm before finding a fixed-point procedure which is fast and accurate. Based on that we introduced the multivariate spatial sign correlation which turned out to be very efficient in high dimensions. First we wanted to simulate the sensivity curve to evaluate its robustness, but results indicated a very simple relationship which arouse my interest and I could indeed calculate the influence function for a special case by using the inverse function theorem. These results can be found in Dürre, Fried and Vogel (2017).

The *third chapter* deals with a robust test for a location change in panel data under serial dependence. Robustness is achieved by using robust scores, which are calculated by applying $\Psi-$functions. The main focus here is to derive asymptotics under the null hypothesis of a stationary panel, if both the number of individuals and time points tend to infinity. We can show under some regularity assumptions that the limiting distribution does not depend on the underlying distribution of the panel as long as we have short range dependence in the time dimension and independence in the cross sectional dimension.

The work was mainly motivated by the article Horváth and Hušková (2012) and extends their work in several directions. I originally studied this article to learn how to prove asymptotics for change-point procedures in the panel context. We wanted to find an easy robustification and hoped that only slight changes in the proofs would be necessary. Actually it turned out that one needs quite different dependence assumptions and only the ideas of the proof could be adopted. Furthermore, the use of $\Psi-$functions necessitates additional rather technical calculations for nuisance parameters. The article has been submitted to the Journal of Time Series Analysis. The preprint Dürre and Fried (2016) is available on arxiv.org.

Chapter 2

Robust estimation of the autocorrelation function

The autocorrelation function (acf) and the partial autocorrelation function (pacf) are elementary tools of linear time series analysis. The graphical presentation as a correlogram gives an idea of the linear dependencies between pairs of observations in different time lags. A sinusoidal shape indicates a seasonality, whereas a slow decay suggests possible long range dependence or non-stationarity. Besides descriptive purposes, the autocorrelation and the autocovariance function can be used for model identification (see Box et al., 1994, pp. 184–188), for fitting autoregressive models using the Yule-Walker equations, for determining the periodogram (see e.g. Brockwell and Davis, 2006; Wei, 1990, pp. 234–238 respectively pp. 265–267), for detecting periodicities (Vecchia and Ballerini, 1991), for clustering or classifying time series (Caiado et al., 2006), and for predicting future values of the time series (Brockwell, 2009).

The sensitivity of the conventional estimators, the sample acf and pacf, to outliers is well known (see Chan, 1992; Deutsch et al., 1990; Maronna et al., 2006, pp. 247–257). A single outlier can drive the sample autocorrelation at every time lag h towards zero, whereas $h+1$ or more successive outliers can make it arbitrarily close to one, both making the estimation worthless. Several robust alternatives have been proposed in the literature to overcome this problem. We review such approaches and evaluate their performances to provide some guidance on which estimator to apply in which data situation.

The first part introduces some notation and background which will be used to describe the robust procedures for estimating the acf and pacf thereafter. It is followed by a simulation study to assess the accuracy and robustness of these estimators. In the final part we draw some conclusions.

2.1 Background and Notation

Let $(X_t)_{t\in\mathbb{Z}}$ denote a real-valued time series. We assume $(X_t)_{t\in\mathbb{Z}}$ to be second order stationary, meaning that the mean and the variance are constant and do not depend on the observation time t, i.e. $\mathrm{E}(X_t) = \mu$ and $\mathrm{Var}(X_t) = \sigma^2 < \infty$ for all $t \in \mathbb{Z}$, while the autocovariance and hence the autocorrelation depend on the time lag only, i.e. $\mathrm{Cov}(X_{t+h}, X_t) = \gamma(h)$ and $\mathrm{Cor}(X_{t+h}, X_t) = \rho(h)$ for all $t, h \in \mathbb{Z}$. Because $\rho(h) = \rho(-h)$, only positive time lags $h \in \mathbb{N}_0$ need to be considered. Both the autocovariance and the autocorrelation functions of a stationary process are always positive-semidefinite, i.e., for every $k \in \mathbb{N}$ the matrix

$$\Gamma^{(k)} = (\Gamma^{(k)}_{i,j})_{i,j=1,\ldots,k+1} \quad \text{with} \quad \Gamma^{(k)}_{i,j} = \gamma(i-j) \tag{2.1}$$

is positive-semidefinite. For a stationary time series the usual relation

$$\mathrm{Cor}(X_{t+h}, X_t) = \frac{\mathrm{Cov}(X_{t+h}, X_t)}{\sqrt{\mathrm{Var}(X_{t+h}) \cdot \mathrm{Var}(X_t)}} \quad \text{implies} \quad \rho(h) = \frac{\gamma(h)}{\gamma(0)}. \tag{2.2}$$

This allows us to translate estimators of the autocovariances into estimators of the autocorrelations and vice versa, if an estimate of the variance $\gamma(0)$ is available.

For a vector of observations $\mathbf{X} = (X_1, \ldots, X_n)$, let $\overline{X}$ be the arithmetic mean, $X_{(1)}, \ldots, X_{(n)}$ denote the ordered sample in ascending order and R_t the rank of X_t, $t = 1, \ldots, n$.

The sample analogues of $\gamma(h)$ and $\rho(h)$ are the empirical or sample autocovariances and autocorrelations $\hat{\gamma}(h)$ and $\hat{\rho}(h)$ (in the simulation study abbreviated as: **Emp. acf**), which are given by

$$\begin{aligned} \hat{\gamma}(h) &= \frac{1}{n}\sum_{t=1}^{n-h}(X_t - \overline{X})(X_{t+h} - \overline{X}), \\ \hat{\rho}(h) &= \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)},\ h \in \mathbb{N}\ . \end{aligned} \tag{2.3}$$

The denominator n is used in the formula for $\hat{\gamma}(h)$ instead of the more intuitive number of cross-products $n - h$, since this guarantees positive-semidefiniteness of the resulting functions $\hat{\gamma}$ and $\hat{\rho}$ for the price of a larger bias. In Schlittgen and Streitberg (2001) p. 244 an asymptotic formula for the bias of the sample acf of Gaussian processes is derived:

$$\mathrm{Bias}(\hat{\rho}(h)) = -\frac{1}{n}\left(h\rho(h) + (1-\rho(h))\sum_{i=-\infty}^{\infty}\rho(i) + 2\zeta(h) - 2\rho(h)\zeta(0)\right) + O(n^{-2}), \tag{2.4}$$

where $\zeta(h) = \sum_{i=-\infty}^{\infty}\rho(i)\rho(i+h)$. Equation (2.4) indicates a large negative bias in case of a small n and a large positive, slowly decaying acf. The estimator is asymptotically unbiased for fixed h as n goes to infinity. The asymptotic distribution of the sample autocorrelation can be found for example in Brockwell and Davis (2006) Theorems 7.2.1 & 7.2.2. Calculation of the empirical acf is recommended only for $n \geq 50$ and $h \leq n/4$ (Box et al., 1994, p. 32).

The sample acf can also be derived from a multivariate covariance estimation. This approach has some desirable features when carried out robustly, as will be seen later on. The matrix $\Gamma^{(k)}$ of the first autocovariances (see (2.1)) can be estimated by building a data matrix from the lagged observations. Let $\tilde{X}_t$, $t \in \mathbb{Z}$, denote the centered observations. We use the sample mean $\overline{X}$ for centering, if not stated otherwise. Defining

$$\boldsymbol{Z}_k' = \begin{bmatrix} \tilde{X}_1 & \tilde{X}_2 & \cdots & \tilde{X}_{k+1} & \cdots & \tilde{X}_n & 0 & \cdots & 0 \\ 0 & \tilde{X}_1 & \cdots & \tilde{X}_k & \cdots & \tilde{X}_{n-1} & \tilde{X}_n & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots & & \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & \tilde{X}_1 & \cdots & \tilde{X}_{n-k} & \tilde{X}_{n-k+1} & \cdots & \tilde{X}_n \end{bmatrix} \in \mathbb{R}^{(k+1)\times(n+k)}, \tag{2.5}$$

the ordinary positive-semidefinite sample autocovariance matrix is obtained from Pearson's product moment covariance estimator

$$\hat{\Gamma}^{(k)} = \mathbf{Z}_k'\mathbf{Z}_k/n. \tag{2.6}$$

Application of the well known identity for correlation matrices,

$$\Xi_{i,j}^{(k)} = \Gamma_{i,j}^{(k)}/\sqrt{\Gamma_{i,i}^{(k)} \cdot \Gamma_{j,j}^{(k)}}, \tag{2.7}$$

yields the estimation $\hat{\Xi}^{(k)}$.

A model for stationary autocorrelation functions is the autoregressive moving average (ARMA) process, which is defined by

$$X_t = \phi_0 + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{i=1}^{q} \theta_i a_{t-i} + a_t, \tag{2.8}$$

with parameters $\phi_0, \phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q \in \mathbb{R}$, and innovations $(a_t)_{t\in\mathbb{Z}}$ forming white noise, that is a stationary sequence of uncorrelated random variables with mean zero and variance σ^2. Of special interest are AR processes where $q = 0$ (Brockwell, 2011), since from them another identity for ρ can be derived. If all solutions z of $1 - \phi_1 z - \ldots - \phi_p z^p = 0$ are outside the complex unit circle, then (2.8) models a stationary process with marginal mean

$$\mu = \frac{\phi_0}{1 - \phi_1 - \ldots - \phi_p}. \tag{2.9}$$

The Yule-Walker equations relate the coefficients $\phi_1, \dots, \phi_p$ of an AR(p) model to the first p autocorrelations $\rho(1), \dots, \rho(p)$. To shorten notation we assume $\phi_0 = 0$. Then the equations are obtained by multiplying (2.8) with X_{t-h}, $h = 1, \dots, p$, taking expectations and dividing by $\gamma(0)$,

$$\begin{aligned} \rho(1) &= \phi_1 + \phi_2\rho(1) + \ldots + \phi_p\rho(p-1) \\ \rho(2) &= \phi_1\rho(1) + \phi_2 + \ldots + \phi_p\rho(p-2) \\ &\vdots \\ \rho(p) &= \phi_1\rho(p-1) + \phi_2\rho(p-2) + \ldots + \phi_p. \end{aligned} \tag{2.10}$$

Autocorrelations of higher order can be extrapolated using the recursion

$$\rho(h) = \phi_1\rho(h-1) + \phi_2\rho(h-2) + \ldots + \phi_p\rho(h-p), \quad h = p+1, p+2, \ldots \tag{2.11}$$

Even if (X_t) is not an AR process of order p, fitting such a model can still be beneficial. Let $\pi_{p,0} + \sum_{i=1}^{p} \pi_{p,i} X_{t-i}$ denote the best approximation of X_t by an AR(p) model in the sense of mean squared error for any $p \in \mathbb{N}$. Then

$$\hat{X}_t = \pi_{h-1,0} + \sum_{i=1}^{h-1} \pi_{h-1,i} X_{t-i} \tag{2.12}$$

is the best linear prediction of X_t given the past and analogously

$$\hat{X}_{t-h} = \pi_{h-1,0} + \sum_{i=1}^{h-1} \pi_{h-1,i} X_{t-h+i} \tag{2.13}$$

is the best linear prediction of X_{t-h} given the future up to time t. The resulting residuals

$$U_{h,t} = X_t - \hat{X}_t \quad \text{and} \quad V_{h,t} = X_{t-h} - \hat{X}_{t-h} \tag{2.14}$$

are called forward respectively backward residuals. They define the partial autocorrelation function (pacf)

$$\pi(h) = \pi_{h,h} = \begin{cases} \operatorname{Cor}(X_{t+1}, X_t), & h = 1 \\ \operatorname{Cor}(U_{h,t}, V_{h,t}), & h \geq 2 \end{cases}, \tag{2.15}$$

which is another important tool for model building. It measures the correlation of X_t and X_{t+h} after eliminating the linear effects of all intermediate observations $X_{t+1}, \dots, X_{t+h-1}$. Unlike the acf, the pacf only needs to be bounded between -1 and 1 to be valid (Ramsey, 1974).

A connection between the acf and pacf is given by the Durbin-Levinson algorithm. For a stationary process with $\mu = 0$, $\gamma(0) > 0$ and $\gamma(h) \to 0$ for $h \to \infty$ it reads

$$\pi(h) = \left(\rho(h) - \sum_{i=1}^{h-1} \pi_{h-1,i}\rho(h-i) \right) v_{h-1}^{-1}, \quad h \geq 2,$$

$$\text{where } \begin{pmatrix} \pi_{h,1} \\ \vdots \\ \pi_{h,h-1} \end{pmatrix} = \begin{pmatrix} \pi_{h-1,1} \\ \vdots \\ \pi_{h-1,h-1} \end{pmatrix} - \pi(h) \begin{pmatrix} \pi_{h-1,h-1} \\ \vdots \\ \pi_{h-1,1} \end{pmatrix}$$

$$\text{and } v_h = v_{h-1}(1 - \pi(h)^2), \tag{2.16}$$

with $\pi_{h,h} = \pi(h)$. The recursion starts with $\pi(1) = \rho(1)$ and $v_0 = 1$. Conversely, the acf can be derived from the pacf using the relationship given by Masarotto (1987)

$$\rho(h) = \sum_{i=1}^{h-1} \pi_{h-1,i}\rho(h-i) + \pi(h) \left(1 - \sum_{i=1}^{h-1} \pi_{h-1,i}\rho(i) \right). \tag{2.17}$$

Instead of estimating the partial autocorrelations (2.15) from the sample acf, Burg proposed an alternative estimator (see Makhoul, 1981) for $\pi(h)$ as

$$\hat{\pi}(h) = 2\frac{\sum_{t=h+1}^{n} U_{h,t}V_{h,t}}{\sum_{t=h+1}^{n}[U_{h,t}^2 + V_{h,t}^2]}. \tag{2.18}$$

It can be interpreted as a correlation estimator for the forward and backward residuals as the denominator estimates the sum of their variances.

In summary, the above equations allow construction of (robust) autocorrelation estimators by estimating ρ either directly, or by estimating the pacf π and using (2.17), or by fitting an AR model of sufficiently large order p and applying (2.10) and (2.11).

2.2 Robust autocorrelation estimators

Different proposals for robust estimation of autocorrelations and partial autocorrelations have been derived using different ideas. We review such approaches in the following.

2.2.1 Estimation based on univariate transformations

An intuitive idea of limiting the influence of outliers is rejecting or at least downweighting very large and small values of the time series, where outlyingness will be relative to the marginal distribution of X_t, ignoring the serial dependence. Such transformations reduce the effects of outliers on the sample acf, but produce a bias which does not vanish asymptotically. An exact bias correction is often not available, so we need to rely on asymptotic approximations or simulations for this. For more details see the section on implementation.

A robust estimator of autocovariances and autocorrelations can be constructed using univariate trimming (abbr.: **Trim**), that is omitting terms in the sum in (2.3) which correspond to the most extreme observations,

$$\hat{\gamma}^{(\alpha)}(h) = \frac{1}{\sum_{t=1}^{n-h} L_t^{(\alpha)} L_{t+h}^{(\alpha)}} \left\{ \sum_{t=1}^{n-h} \left(X_t - \bar{X}^{(\alpha)} \right) \left(X_{t+h} - \bar{X}^{(\alpha)} \right) L_t^{(\alpha)} L_{t+h}^{(\alpha)} \right\},$$

$$\text{where } \bar{X}^{(\alpha)} = \frac{1}{\sum_{t=1}^{n} L_t^{(\alpha)}} \sum_{t=1}^{n} X_t L_t^{(\alpha)}$$

$$\text{and } L_t^{(\alpha)} = \begin{cases} 1, & X_{(g)} < X_t < X_{(n-g+1)} \\ 0, & \text{otherwise} \end{cases} \text{ with } g = \lfloor \alpha \cdot n \rfloor \text{ for } 0 \leq \alpha < 0.5.$$

Chan and Wei (1992) propose trimming constants α between 0.01 and 0.1, depending on the suspected percentage of outliers. As usually, larger fractions α increase robustness but decrease the efficiency of the estimator at clean samples without outliers. The acf is estimated by dividing the trimmed autocovariance through the trimmed variance $\hat{\gamma}^{(\alpha)}(0)$. Simulations indicate that without a bias correction the estimator is significantly biased for $n = 50$ (see also Chan and Wei, 1992), and it may be easily seen that the bias does not vanish asymptotically if α is fixed.

To obtain high robustness, Chakhchoukh (2010) suggests substituting the sum in the sample acf by the median, calculating

$$\hat{\rho}(h) = \frac{\operatorname{med}(\tilde{X}_1\tilde{X}_{1+h}, \ldots, \tilde{X}_{n-h}\tilde{X}_n)}{\operatorname{med}(\tilde{X}_1^2, \ldots, \tilde{X}_n^2)},$$

where $\tilde{X}_t$ is the centered time series, for example using the median. This estimator (abbr.: **Mediancor**) can be seen as a limiting case of the above trimming based estimator, with $\alpha = 0.5$. For an asymptotically consistent estimation of $\rho(h)$ a nonlinear transformation of $\hat{\rho}(h)$ is necessary, which needs to be determined numerically.

With the aim of robustly fitting time series models, Bustos and Yohai (1986) introduces the so called residual autocovariances (RA-estimators), which can also be used to estimate the acf. Albeit being defined more generally, this approach boils down to a more sophisticated transformation of the time series (see Bustos and Yohai, 1986, for the general definition). Instead of trimming a constant amount of the largest and smallest observations, observations are downweighted only if being unusually large or small. Note that the amount of rejected observations depends on the sample itself. For the transformed time series Y_t, $t = 1, \ldots, n$, one gets

$$Y_t = \psi\left(\frac{X_t - m}{s}\right) \tag{2.19}$$

where m and s are suitable estimators for μ and $\sqrt{\gamma(0)}$. The median and the median absolute deviation about the median (MAD) are common robust choices for these quantities. Conventional choices of the transformation function ψ with tuning parameters c_j are the Huber function

$$\psi(x) = \psi_{c_1}(x) = \operatorname{sign}(x)\min(|x|, c_1) \tag{2.20}$$

and Tukey's bisquare function

$$\psi(x) = \psi_{c_2}(x) = \begin{cases} x(1 - x^2/c_2^2)^2, & 0 \leq |x| \leq c_2 \\ 0, & |x| > c_2. \end{cases} \tag{2.21}$$

The resulting estimators are abbreviated by **RA-Huber** and **RA-Tukey**. The objective of Bustos and Yohai (1986) was not estimation of the acf, so a bias correction was not proposed. However, tuning parameters like $c_1 = 1.37$ for the Huber function and $c_2 = 4.68$ for Tukey's function modify a Gaussian time series only slightly in the absence of outliers, so that the resulting bias is small.

2.2.2 Estimation based on signs and ranks

For the purpose of model selection Garel and Hallin (1999) introduces rank based statistics, which can also be applied for acf estimation. Construction of ranks means a special data transformation as treated in the previous subsection. Nevertheless we present this approach separately together with sign based estimators since both are popular utilities from nonparametric statistics and often mentioned together. Additionally, bias corrections are known explicitly at least for Gaussian processes.

Since we are more interested in estimation than in testing, we use a definition slightly different from Garel and Hallin (1999), namely

$$\hat{\rho}(h) = c \sum_{i=1}^{n-h} J(R_i/(n+1)) \cdot J(R_{i+h}/(n+1)) \tag{2.22}$$

with $c = 1/\sum_{i=1}^{n} J(R_i/(n+1))^2$ and J a score function. Van der Waerden or normal scores are obtained by

$$J(x) = \Phi^{-1}(x), \quad x \in (0,1),$$

where $\Phi(x)$ is the cumulative distribution function of a standard normal, and lead to asymptotically optimal tests under normality (Garel and Hallin, 1999). The asymptotical Gaussian efficiency of the resulting estimator is higher than those of other rank based estimators (Ferretti et al., 1991). However, the related Gaussian rank correlation (abbr.: **GRCor**) is not very robust against outliers (Boudt et al., 2012).

More widely used is the Spearman score function $J(x) = x - (n+1)/2$, which result in an autocorrelation estimator based on the popular Spearman's ρ (abbr.: **Spearman**). Whereas van der Waerden scores yield an asymptotically unbiased estimation in the normal case, Spearman's ρ needs to be transformed by $g(\rho) = 2\sin(\pi\rho/6)$, see for example Croux and Dehon (2010).

Further popular nonparametric correlation estimators are Kendall's τ (abbr.: **Kendall**)

$$\hat{\rho}(h) = \frac{2}{(n-h)(n-h-1)} \sum_{i>j} \operatorname{sign}\left((X_i - X_j)(X_{i+h} - X_{j+h})\right)$$

and the quadrant correlation (abbr.: **Quadrant**)

$$\hat{\rho}(h) = \frac{1}{n-h} \sum_{i=1}^{n-h} \operatorname{sign}\left((X_i - \hat{\mu})(X_{i+h} - \hat{\mu})\right),$$

where the center $\hat{\mu}$ is estimated by the median of the time series. For both estimators transformation by $g(\rho) = \sin(\pi\rho/2)$ yields unbiasedness under the bivariate normal distribution, and also for a wider range of distributions (Möttönen et al., 1999). A disadvantage of such transformations is that they can destroy the positive-semidefiniteness of the estimators.

2.2.3 Estimation based on partial autocorrelation

Autocorrelation estimators constructed from pairwise correlation estimators possibly lack positive-semidefiniteness as mentioned before. Valid estimation of the pacf is easier, since one only needs to ensure estimates within -1 and 1. Using relation (2.17) between pacf and acf with initialization $\hat{\pi}(1) = \hat{\rho}(1)$ then yields a positive semidefinite autocorrelation estimation. This motivates estimating the pacf first, as suggested by Masarotto (1987) and Möttönen et al. (1999). Both approaches differ in the choice of the correlation estimator for $\pi(h)$ based on the identity (2.15).

An M-estimator as a variant of the Burg estimator (2.18) is proposed by Masarotto (1987) and defined as the solution $\hat{\pi}(h)$ of

$$\sum_{t=h+1}^{n} W_{h,t}(d_{ht}(\hat{\pi}(h))/s_{hn}^2)\left(2U_{ht}V_{ht} - \hat{\pi}(h)(U_{ht}^2 + V_{ht}^2)\right), \tag{2.23}$$

where $W_{h,t} = w(d_{ht}(b)/s_{ht}^2)$ with weight function $w(x) = 3/(1+x)$, $d_{ht}(b) = U_{ht}^2 + V_{ht}^2 - 2bU_{ht}V_{ht}$ and s_{ht} satisfying

$$\sum_{t=h+1}^{n} w(d_{ht}(b)/s_{ht}^2)d_{ht}(b) = 2(n-h)s_{ht}^2.$$

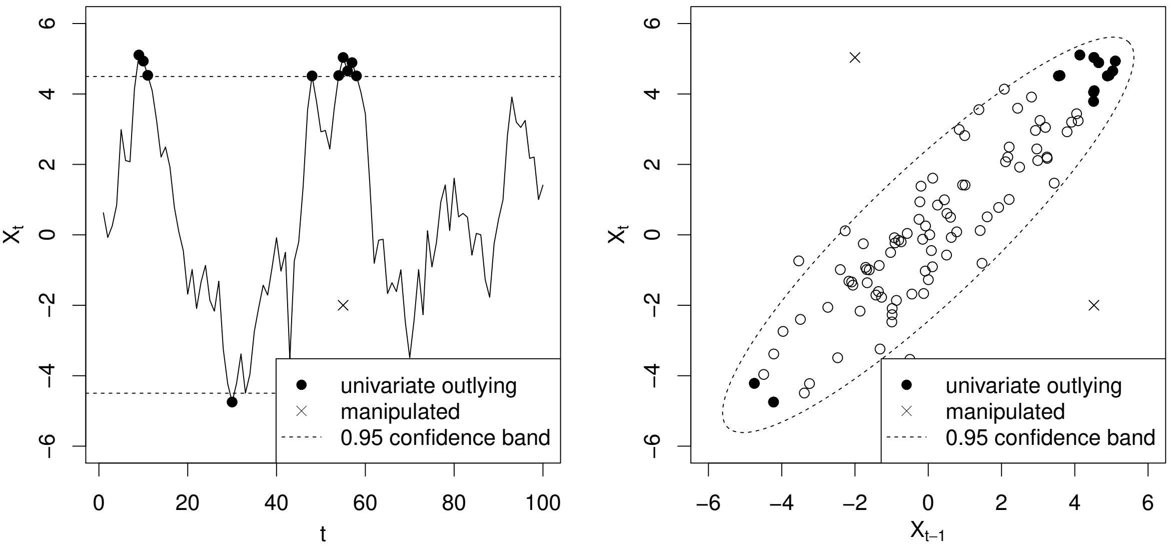

Figure 2.1: Time series with 95% prediction bounds based on the univariate (left) and the bivariate (right) marginal distribution corresponding to subsequent observations. Univariate margins identify the most extreme observations as outliers, while multivariate inspection takes the dependencies between subsequent observations into account and identifies the true outliers.

Masarotto (1987) argues that the resulting acf estimator (abbr.: **PA-M**) is consistent and asymptotically normal at least under normality. Asymptotical confidence bands can be constructed numerically. As an alternative, Möttönen et al. (1999) proposes sign and rank based correlation estimators, e.g. Spearman's ρ (abbr.: **PA-Spearman**), Kendall's τ (abbr.: **PA-Kendall**) and quadrant-correlation (abbr.: **PA-Quadrant**), as described in the previous subsection. Generally, every robust bivariate correlation estimator can be applied.

2.2.4 Estimation based on multivariate correlation

Approaches based on univariate transformations ignore the serial dependence of the data, possibly downweight good observations and overlook outliers. The left panel of Figure 2.1 depicts a realization of an AR process with $\phi_0 = 0$ and $\phi_1 = 0.9$. Prediction bounds based on the univariate marginal distribution simply identify the most extreme observations as possible outliers, although these observations might be due to the dynamics of the underlying process. A bivariate or even multivariate analysis based on the marginal distribution of subsequent observations (Gather et al., 2002) allows us to take the dependencies among subsequent observations into account, and can achieve better downweighting of spurious observations in the subsequent analysis than a simple univariate consideration. Estimation of the acf from a robust estimate of the multivariate covariance matrix is thus promising. Such estimators can be based e.g. on multivariate trimming or weighting. Moreover, some multivariate robust correlation estimators even gain efficiency with increasing dimension (Taskinen et al., 2006).

Multivariate methods can be formulated in terms of the data matrix $\mathbf{Z}_k$ in (2.5). Note that centering is unnecessary, since the described approaches estimate a robust center. The computing time of many robust procedures increases exponentially in the dimension (Vakili and Schmitt, 2014), so one should choose k rather small. To simplify notation, we denote the i-th row of $\mathbf{Z}_k$ as $\mathbf{M}_i'$, so that we are in the usual multivariate case. The estimation result will always be a valid covariance matrix but it does not have the Toeplitz structure with constant off-diagonals, albeit by definition all values of the h-th off-diagonal estimate $\rho(h)$. An intuitive solution is averaging the values across each off-diagonal, i.e.

$$\hat{\rho}(h) = \frac{1}{k-h+1} \sum_{i=1}^{k-h+1} \hat{\Xi}^{(k)}_{i,i+h}, \tag{2.24}$$

though positive-semidefiniteness gets possibly lost then.

There is a large literature on robust multivariate correlation estimation. We will concentrate on the most common proposals, but of course others could be employed as well.

An M-estimator of scatter (abbr.: **Multi-M**), which can be represented as a weighted least squares estimate, is introduced in Maronna (1976), see also Maronna et al. (2006). Given an initial estimator $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ for expectation and covariance, robust weights are obtained from the outlyingness of the observations as measured by the Mahalanobis distance

$$d_i^2 = (\mathbf{M}_i - \hat{\boldsymbol{\mu}})'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{M}_i - \hat{\boldsymbol{\mu}}), i = 1, \ldots, n+k. \tag{2.25}$$

After that, the estimation is sequentially updated by

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n+k} v(d_i)(\mathbf{M}_i - \hat{\boldsymbol{\mu}})(\mathbf{M}_i - \hat{\boldsymbol{\mu}})' \quad \text{and} \quad \hat{\boldsymbol{\mu}} = \frac{\sum_{i=1}^{n+k} w(d_i)\mathbf{M}_i}{\sum_{i=1}^{n+k} w(d_i)}\ , \tag{2.26}$$

where $w(d_i)$ and $v(d_i)$ are suitable weights. Popular are Huber weights: $w(d_i) = \min(1, c_0/|d_i|)$ and $v(d_i) = w(d_i)^2/r$, where c_0 determines robustness and efficiency and r depending on c_0 and the probability model ensures consistency. Using the weight function $v(d) = (k+1)/d^2$ results in Tyler's M-estimator (abbr.: **Multi-TylerM**), which is a kind of minimax estimator within the elliptical model (Tyler, 1987).

The breakdown point (see Hubert and Debruyne, 2009, for definition and meaning) of M-estimators cannot exceed an upper bound which decreases with increasing dimension (Maronna et al., 2006). Since the effective amount of outlying vectors in the estimation of the acf can be k-times the number of outlying observations, other estimators might be preferred if one is interested in larger time lags.

The disadvantage of the decreasing breakdown point does not apply to multivariate S-estimators (Davies, 1987) (abbr.: **Multi-S**). They are defined as

$$\hat{\boldsymbol{\Sigma}} = \underset{\boldsymbol{\mu},\, \boldsymbol{\Sigma}}{\arg\min} \left\{ \det(\boldsymbol{\Sigma}) : \frac{1}{n+k} \sum_{i=1}^{n+k} w\left((\mathbf{M}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{M}_i - \boldsymbol{\mu})\right) = b_0 \right\},$$

where w is a bounded smooth and nondecreasing function, e.g.

$$w(y) = \min\left(\frac{y^2}{2} - \frac{y^4}{2c^2} + \frac{y^6}{6c^4}, \frac{c^2}{6}\right).$$

The constant c determines the breakdown point, whereas b_0 depends on the probability model; see Lopuhaä (1989) for more details. An algorithm for computing this implicitly defined estimator can be found in Salibian-Barrera and Yohai (2006). Although the breakdown point does not depend on the number of dimensions, single outliers can cause a larger bias in higher dimensions (Maronna et al., 2006).

A popular robust covariance estimator is the minimum covariance determinant (abbr.: **Multi-rMCD**) (Rousseeuw, 1985; Hubert and Debruyne, 2010). For a given constant α between 0 and 0.5 the usual product moment covariance is calculated for the subset of proportion $1 - \alpha$ which leads to the matrix with the smallest determinant. An approximate procedure was proposed by Rousseeuw and Driessen (1999), since finding this subset is very time consuming for large n. Larger trimming constants α lead to more robust but less efficient estimators, with the efficiency for large α being rather low (Croux and Haesbroeck, 1999). To combine high robustness and large efficiency, often an additional reweighting step is added (abbr.: **Multi-wMCD**): Robust Mahalanobis distances are obtained based on an initial MCD fit, and then the ordinary covariance matrix is calculated from all observations with Mahalanobis distances not exceeding a certain quantile of the χ^2-distribution with $k+1$ degrees of freedom. The 0.975-quantile has been recommended for this cut-off (Rousseeuw and Driessen, 1999). An asymptotically fully efficient reweighting step (abbr.: **Multi-effMCD**) with a data-adaptive choice of the quantile was suggested in Gervini (2003).

Multivariate outliers can be inconspicuous if one only looks at individual dimensions, but there is always a one-dimensional projection in which the observation is outlying (Hadi et al., 2009). Based on this idea, Stahel (1981) and Donoho (1982) propose to use the maximal distance to the median for every possible projection to measure outlyingness, i.e.

$$r_i = \max_{\boldsymbol{a}:\|\boldsymbol{a}\|=1} \frac{\boldsymbol{a}'\mathbf{M}_i - \text{Median}(\boldsymbol{a}'\mathbf{M}_1, \ldots, \boldsymbol{a}'\mathbf{M}_{n+k})}{\text{MAD}(\boldsymbol{a}'\mathbf{M}_1, \ldots, \boldsymbol{a}'\mathbf{M}_{n+k})},$$

with $\|\cdot\|$ being the Euclidean norm. Practical algorithms only consider a finite set of randomly chosen vectors for $\boldsymbol{a}$. The number of such directions needs to increase strongly for higher dimensions to ensure reliable outlier detection. The resulting Stahel-Donoho estimator (abbr.: **Multi-SD**) is defined as the weighted covariance

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n w_i(\mathbf{M}_i - \hat{\boldsymbol{\mu}})(\mathbf{M}_i - \hat{\boldsymbol{\mu}})' \quad \text{and} \quad \hat{\boldsymbol{\mu}} = \frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n w_i \mathbf{M}_i.$$

A common choice of the weight function is $w_i = \min\left(1, (c/r_i)^2\right)$, and c is often chosen as the 0.975-quantile of the χ^2-distribution with $k+1$ degrees of freedom (Croux and Haesbroeck, 1999).

2.2.5 Estimation based on variances

An estimation principle for covariances and correlations based on estimators of variances has been proposed in Gnanadesikan and Kettenring (1972). In the context of autocorrelation estimation for stationary time series, the underlying formula reads

$$\rho(h) = \text{Cor}(X_{t+h}, X_t) = \frac{\text{Var}(X_{t+h} + X_t) - \text{Var}(X_{t+h} - X_t)}{\text{Var}(X_{t+h} + X_t) + \text{Var}(X_{t+h} - X_t)}, \tag{2.27}$$

see Ma and Genton (2000). The usual correction factors necessary for making robust scale estimators consistent at a certain distribution are not needed when applying them for correlation estimation, since they cancel out if $X_{t+h} + X_t$ and $X_{t+h} - X_t$ are in the same location-scale family. This is fulfilled, e.g., if X_{t+h} and X_t are jointly normal or, more generally, elliptically-symmetric distributed. Note that this approach does not necessarily yield a positive-semidefinite estimation of the acf.

For estimation of the variances on the right hand side of (2.27), trimmed and winsorized variances have been suggested (Gnanadesikan and Kettenring, 1972). Since any reasonable estimator of variability can be applied, Ma and Genton (2000) propose Q_n (Rousseeuw and Croux, 1993) (abbr.: **GK-Qn**), because of its high asymptotic breakdown point of 0.5 and its good asymptotic efficiency of 0.82 for i.i.d. Gaussian samples. The Q_n corresponds roughly to the first quartile of all absolute pairwise distances between all pairs of observations.

In the context of ordinary correlation Maronna and Zamar (2002) recommended the so called τ-scale estimator (abbr.: **GK-tau**)

$$\hat{\sigma}^2(X_1, \ldots, X_n) = \frac{\hat{\sigma}_0^2}{n} \sum_{i=1}^n d_{c_2}\left(\frac{X_i - \hat{\mu}}{\hat{\sigma}_0}\right), \tag{2.28}$$

where $\hat{\mu}$ is an adaptively weighted mean of the observations, $\hat{\sigma}_0$ their MAD and $d_c(x) = \min(x^2, c^2)$. Tuning constants of $c_1 = 4.5$ (for $\hat{\mu}$) and $c_2 = 3$ results in a good trade-off between efficiency and robustness and an asymptotic Gaussian efficiency of 0.8 in case of independent observations (Maronna and Zamar, 2002). The good properties of this estimator in the bivariate i.i.d. case are promising for the estimation of autocorrelations.

2.2.6 Estimation based on robust filtering

As mentioned above, clean observations can be outlying with respect to the marginal distribution and thus be unnecessarily downweighted by estimators based on univariate transformations, if the autocorrelations $\rho(h)$ are large positive and slowly decaying. The robust filtering approach overcomes this problem by taking the time series structure into account. The idea is to measure the outlyingness of the prediction residuals $U_{p,t}$ instead of X_t itself. After replacing outliers by reasonable values, one can either calculate the sample acf (abbr.: **Filter-acf**) or use the fitted AR process and translate this into the acf via the Yule-Walker equations (abbr.: **Filter-ar**). Robust filtering was already introduced by Masreliez (1975), but we will stick to the filter described in Maronna et al. (2006), which is a slight modification proposed by Martin and Thomson (1982). Note that this algorithm is quite extensive so we will summarize only the main ideas and refer to Maronna et al. (2006, pp. 272–277 and 320–321) for details.

Let $\tilde{X}_t$ be centered for example by the median and approximate the process by an AR model of order $p \in \mathbb{N}$. A kind of robust AIC criterion to determine p was proposed by Maronna et al. (2006).
Let $\mathbf{Y}_t = (Y_t, \ldots, Y_{t-p+1})'$ denote the vector of robustly filtered values and

$$\Phi = \begin{pmatrix} \phi_1, \ldots \phi_{p-1} & \phi_p \\ \mathbf{I}_{p-1} & \mathbf{0}_{p-1} \end{pmatrix} \tag{2.29}$$

the so called transition matrix. From this one calculates the one step ahead predictions

$$\hat{X}_t = \sum_{i=1}^{p} \phi_i Y_{t-i}$$

and its residuals

$$\tilde{U}_t = X_t - \hat{X}_t.$$

Note that this is similar to usual prediction residuals defined in (2.14), just replacing X_j by Y_j for $j = t-1, \ldots, t-p$ to make it more robust. Eventually one sets

$$\mathbf{Y}_t = \Phi \mathbf{Y}_{t-1} + \frac{\boldsymbol{m}_t}{s_t} \psi \left(\frac{\tilde{U}_t}{s_t} \right), \tag{2.30}$$

where ψ should fulfill

$$\psi(x) = \begin{cases} x & |x| < d_1 \\ 0 & |x| > d_2, \end{cases} \tag{2.31}$$

with $0 < d_1 < d_2$. Our proposal is to use a polynomial of degree three between d_1 and d_2. It is uniquely defined by forcing ψ to be continuous differentiable:

$$\psi(x) = |x|(a + b|x| + dx^2 + e|x|^3) \quad \text{for} \quad d_1 \leq |x| \leq d_2, \tag{2.32}$$

where

$$a = \frac{2d_1^2 d_2^2}{(d_1 - d_2)^2} \qquad b = \frac{-d_2^3 - d_1 d_2^2 - 4d_1^2 d_2}{(d_1 - d_2)^2}$$

$$d = \frac{2d_2^2 + 2d_1 d_2 + 2d_1^2}{(d_1 - d_2)^2} \qquad e = \frac{-d_1 - d_2}{(d_1 - d_2)^3}.$$

Furthermore $\boldsymbol{m}_t \in \mathbb{R}^p$ contains estimations of the variance of the prediction residual s_t^2 and the covariances between the residual $\tilde{U}_t$ and the robustly filtered values $Y_{t-1}, \ldots Y_{t-p+1}$. Note that these estimations are time-dependent instead of a simpler global estimation. The reason for this is that in case of outliers there is a chance that the algorithm otherwise might lose track of the data afterwards. In this case s_t will increase and thus provide the filtered values more variation to get back to the data more quickly. See Martin and Thomson (1982) for more details. Recursions for $\boldsymbol{m}_t$ are given as

$$\boldsymbol{M}_t = \Phi \boldsymbol{P}_{t-1} \Phi' + \boldsymbol{d}\boldsymbol{d}' \hat{\sigma}_u^2$$

and

$$\boldsymbol{P}_t = \boldsymbol{M}_t - \frac{1}{s_t^2} \psi \left(\frac{\tilde{U}_t}{s_t} \right) \frac{s_t}{\tilde{U}_t} \boldsymbol{m}_t \boldsymbol{m}_t'$$

where $\boldsymbol{m}_t$ is the first column of $\boldsymbol{M}_t$, $\boldsymbol{d} = (1, 0, \ldots, 0)' \in \mathbb{R}^p$ and $\hat{\sigma}_u^2$ an estimator for the variance of the prediction residuals.

Looking only at the first row of equations in (2.30) we get

$$Y_t = \hat{X}_t + s_t \psi \left(\frac{\tilde{U}_t}{s_t} \right), \tag{2.33}$$

indicating that Y_t will be close to X_t if $|\tilde{U}_t|$ is small, and close to $\hat{X}_t$ if it is large. Using the vector recursion instead of a simpler one dimensional equation (2.33) offers the advantage that if X_t is an outlier, the algorithm will also use future information on $X_s,\ s > t$, to determine $\hat{X}_t$.

The parameters $\phi_1, \ldots, \phi_p$ can be estimated by minimizing the variance of the prediction residuals

$$\hat{\sigma}(\tilde{U}_{p+1}(\phi_1, \ldots, \phi_p), \ldots, \tilde{U}_n(\phi_1, \ldots, \phi_p)). \tag{2.34}$$

For $\hat{\sigma}$ Maronna et al. (2006) proposed the τ-scale (2.28) because of its quick computation and good robustness. Since a non-convex function needs to be optimized over p parameters, they suggested sequential minimization based on the Durbin-Levinson algorithm. This converts the problem into p one-dimensional optimizations, which are proposed to be optimized by a line search.

2.2.7 Positive-semidefiniteness

From the above approaches the usual sample acf, the procedures using partial autocorrelations, the acf of the robustly filtered values as well as the Gaussian rank autocorrelation are guaranteed to be positive-semidefinite.

Bivariate correlation estimators do not necessarily yield positive-definite correlation matrices unless they calculate the usual correlation based on transformed data. A further problem arises for multivariate correlation estimators, resulting in positive-semidefinite matrices which do not

Estimators	Tuning parameters
RA-Huber	$c_1 = 1.37$
RA-Tukey	$c_2 = 4.68$
Trim	$\alpha = 0.1$
Multi-M	$c_0 = \sqrt{F_{k+1}^{-1}(0.9)}$, $r = F_{k+3}(c_0^2) + \frac{c_0^2 \cdot 0.1}{k+1}$ with $F_k \sim \chi_k^2$
Multi-S	$c = 6.02$ (maximal lag of 7)
Multi-MCD	$\alpha = 0.5$
Multi-SD	$c = \sqrt{F_{k+1}^{-1}(0.95)}$ with $F_k \sim \chi_k^2$
GK-tau	$c_1 = 4.5$, $c_2 = 3$
Filter-acf / ar	$c_1 = 4.5,\ c_2 = 3$ for (2.34), $d_1 = 2,\ d_2 = 3$ for (2.31)

Table 2.1: Tuning parameters used in our simulation study.

possess a Toeplitz structure, meaning that there will be different values on the off-diagonals. Enforcing this by averaging the off-diagonals, for instance, can destroy the positive-semidefiniteness. Construction of the empirical counterpart of the correlation matrix $\Xi^{(k)}$ defined in formula (2.7) allows to apply transformations which achieve positive-semidefiniteness, but this destroys the Toeplitz structure.

A more appealing approach is finding the best positive-semidefinite Toeplitz approximation, minimizing e.g. the Frobenius norm (Al-Homidan, 2006). In our simulations we use the simple projection method proposed there, which can be described as follows. Let A be any real symmetric matrix and $A = UDU'$ denote an eigenvalue decomposition, where D is a diagonal matrix containing all eigenvalues. If A is not positive-semidefinite there will be some eigenvalues smaller than zero. Setting these to zero yields the matrix $\tilde{D}$ and results in a projection $P_p(A) = U\tilde{D}U'$, which is positive-semidefinite but not Toeplitz. For any matrix $B \in \mathbb{R}^{p \times p}$ we denote by $P_T(B)$ the matrix which results from setting all off-diagonal elements of B of order j to its average for $j = 1, \ldots, p-1$. Then the projection method for the best positive-semidefinite Toeplitz approximation can be described as

$$\begin{aligned} R_0 &= \hat{\Xi}^{(k)} \\ R_{i+1} &= R_i + [P_p(P_T(R_i)) - P_T(R_i)] \qquad\qquad i = 1, \ldots \end{aligned}$$

and is stopped if the change in the Frobenius norm $||R_{i+1} - R_i||_F$ becomes negligible, see Al-Homidan (2006).

2.3 Simulations

There are only a few theoretical results available to compare the different autocorrelation estimators. We thus perform a simulation study using the statistical software R (R Core Team, 2016) and the robts package (Dürre et al., 2016). Many of the proposed estimators require the choice of tuning parameters. For the purpose of a fair comparison these parameters are often selected by making all estimators equally efficient in simulation studies. Unfortunately, there are many estimators without tuning possibility and a lack of theoretical results so that this is not possible here. We use the tuning parameters proposed by the respective authors instead. If possible we favor more robust versions, since our simulation scenarios often contain highly contaminated data. The chosen tuning parameters can be found in Table 2.1.

Abbr.	AR[0]	AR[0.4]	AR[0.8]	AR[-0.4]
ϕ_1	0	0.4	0.8	-0.4
θ_1	0	0	0	0

Abbr.	AR[-0.8]	MA[0.4]	ARMA[0.4,0.4]	ARMA[0.8,-0.4]
ϕ_1	-0.8	0	0.4	0.8
θ_1	0	0.4	0.4	-0.4

Table 2.2: Considered processes and their abbreviations.

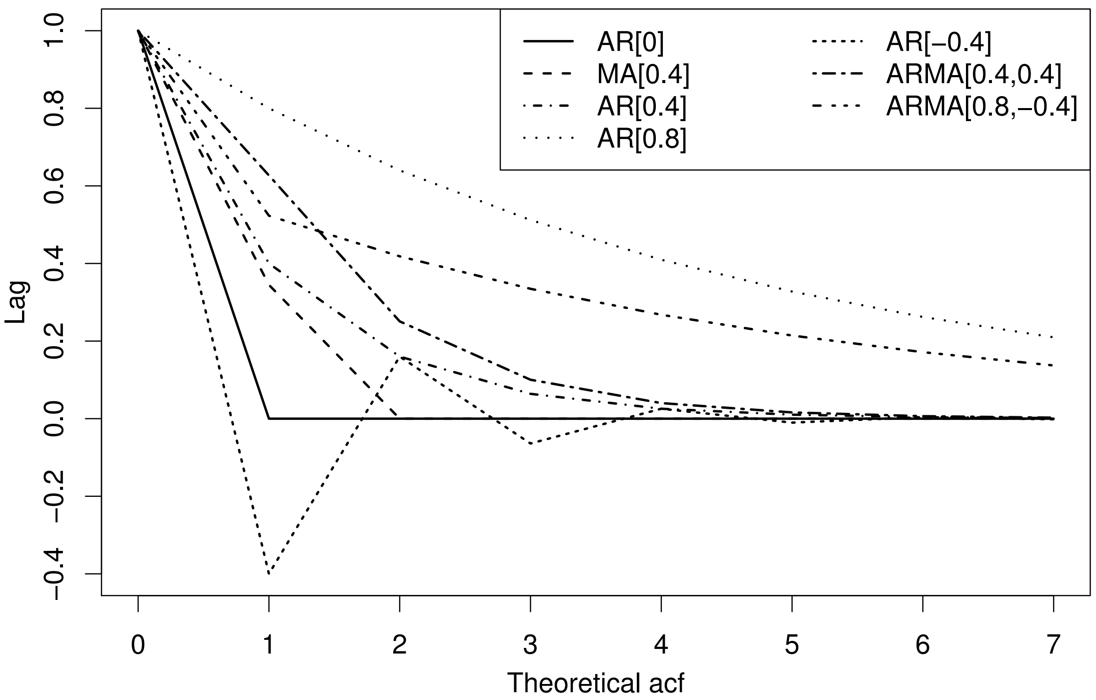

Figure 2.2: Autocorrelation functions of the processes considered in the simulations.

We mainly focus on first order autoregressive moving average (ARMA) processes (as defined in 2.8) because of their simplicity and popularity, considering the seven parameter settings shown in Table 2.2 along with their abbreviations. Figure 2.2 indicates that this selection covers rather different autocorrelation functions. If not explicitly stated otherwise, the innovations are standard normal.

We calculate the acf only for the first seven lags for different reasons. Multivariate correlation estimators are time consuming for large lags and the acf of most of the processes is nearly zero for lags larger than six. So we do not expect qualitatively different behavior for higher time lags. However, simulations indicate a slight loss of efficiency of robust estimators for higher time lags.
To simplify the comparison we consider maximal bias and minimal efficiency across all lags instead of looking at all seven time lags separately. Simulations reveal that the maximal (absolute) bias

$$\max_{h=1,\dots,7} |\mathrm{Bias}(\hat{\rho}(h))|$$

is usually realized for $h = 1$, whereas the minimal efficiency compared to the sample acf $\tilde{\rho}$

$$\min_{h=1,\dots,7} (\mathrm{MSE}(\tilde{\rho}(h))/\mathrm{MSE}(\hat{\rho}(h)))$$

occurs often for $h = 1$ or the largest lag considered here, $h = 7$. In the case of contaminated data, we calculate the efficiency relative to the sample acf for clean data. This measures the amount of information lost due to outliers when using a robust estimator.

2.3.1 Efficiency for uncontaminated data

First we investigate the properties in case of clean data without outliers, starting with the AR[0.4] model. The results are based on 10 000 runs each. As mentioned before, the empirical acf is biased for small n. As can be seen in Figure 2.3, the small sample bias is comparable to that of the robust alternatives. Usually there is a bias towards 0, except for the PA-Quadrant and the robust filtering approach. For the latter the bias changes from negative values for small n to slightly positive values for large n. This is not surprising. It was already mentioned by Maronna et al. (2006) that smoothing the time series produces a nonvanishing bias. For small n this is overruled by the natural negative bias of the autocorrelation estimation. Tyler's M-estimator and RA-estimators are less biased than the other methods for small n, resulting in a good finite sample efficiency. Generally, multivariate S- and M-estimators achieve high efficiencies. GK, rank and sign based approaches and also the reweighted MCD versions need larger samples to get a small MSE. As opposed to this, the relative efficiency of the estimation by the raw MCD, Tyler's M-estimator, the RA-approaches, the Filter-ar method and the median correlation decreases. For the MCD this is in line with simulation results in the multivariate case (see Croux and Haesbroeck, 1999). For the Filter-ar algorithm we notice an increase of the variance relative to the empirical acf, whereas a slower decay of the bias is the reason for the other estimators.

The findings for other models are similar. For processes with strong positive autocorrelations the maximal bias increases for all estimators just as the minimal efficiency. Nevertheless, the order of the estimators with respect to efficiency nearly stays the same. The Filter algorithms are an exception having a relatively low efficiency compared to the other approaches if the process has small absolute correlations. An explanation could be that, generally, the estimator of the fitted AR parameter ϕ_i has a large variance in this case, which affects the precision of the filtered values.

In time series we often face distributions with tails heavier than the Gaussian (Davis and Resnick, 1986; Loretan and Phillips, 1994; Politis, 2009; Rojo, 2013). Estimators should remain reliable in case of such departures from normality. Therefore we considered maximal absolute bias and minimal efficiency for AR models with t-distributed innovations of different degrees of freedom. Already for three degrees of freedom the results were similar to those under normality. Only estimators based on partial autocorrelation lose considerable efficiency. Note that three degrees of freedom correspond to the heaviest tails possible for which the acf is defined under t-distributions. Our simulations agree with the theoretical result that the sample acf of a linear process is still $\sqrt{n}$-consistent without the need of fourth moments, see Davis and Mikosch (1998).

2.3.2 Robustness under contamination

Additive outliers are known to be particularly harmful for the estimation of dependence parameters. Such outliers describe e.g. measurement errors, where a certain value ω is added to the observation at time t, see Fox (1972). While for the empirical acf the effect of an outlier increases in ω due to monotonicity, for robust estimators this influence is generally bounded. However, while the influence is still monotone in ω for rank-based or other monotone estimators, it can

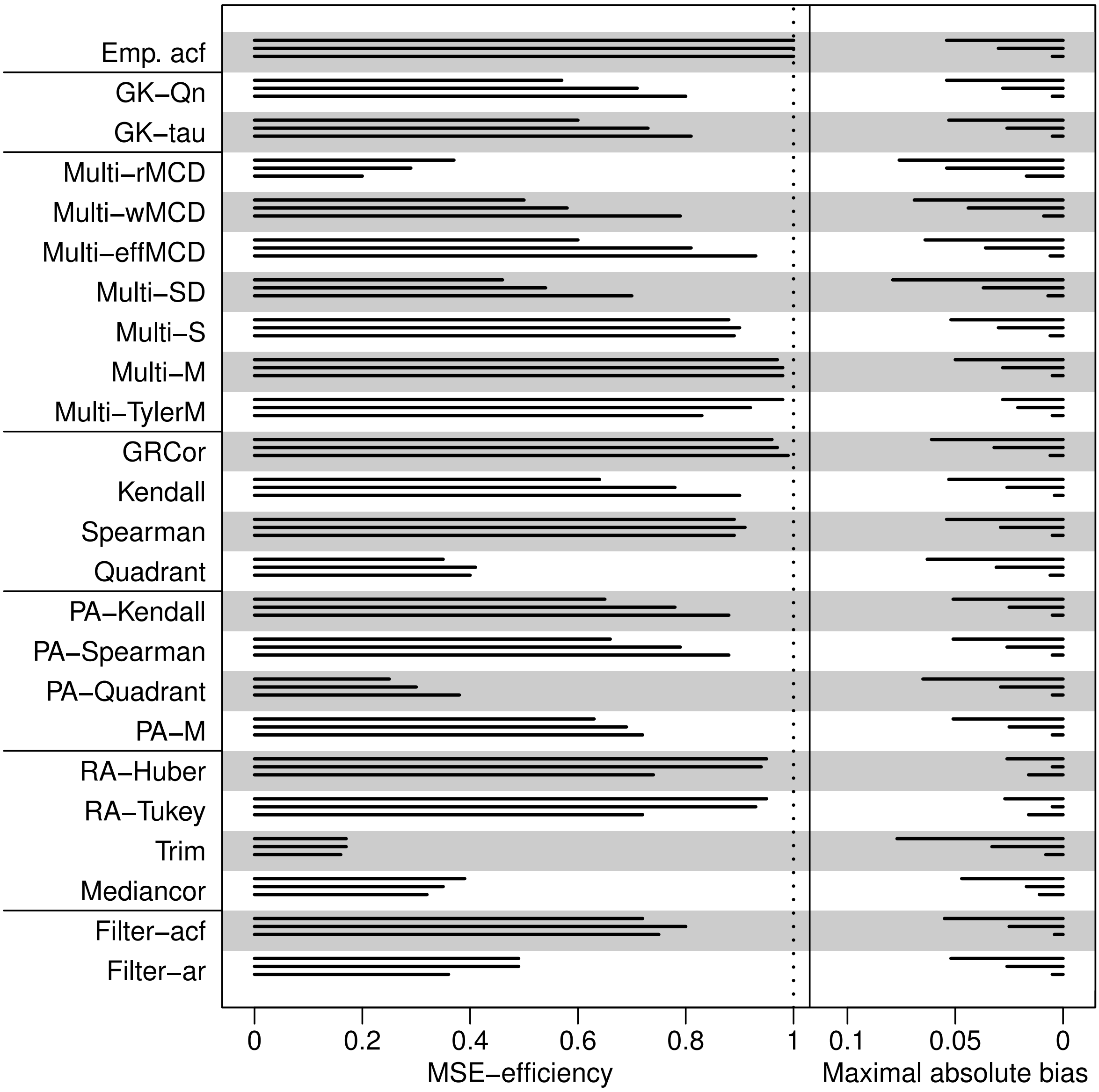

Figure 2.3: Efficiency (left) and bias (right) for $n = 50,\ 100,\ 500$ (from top to bottom in each panel) for an AR[0.4] model with normal innovations.

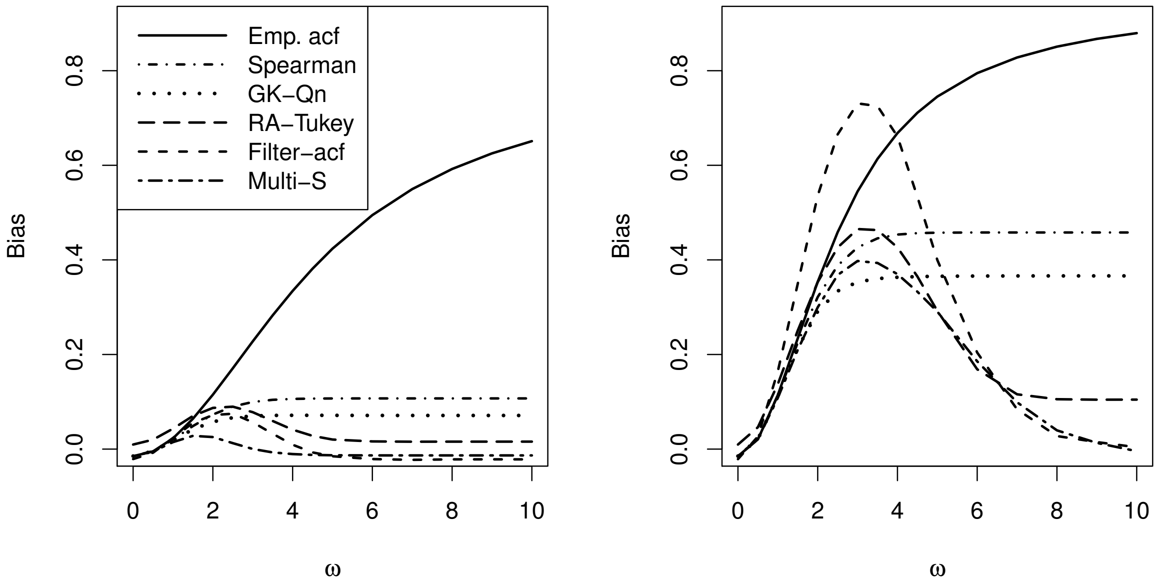

Figure 2.4: Simulated bias of a contaminated AR[0] model with $n = 100$ and a patch of 5 (left) or 20 outliers (right).

even decrease for very large values of ω for other estimators, e.g. for so called redescenders like S-estimators, see Figure 2.4. This means that different outlier sizes are worst-case for different estimators. Since we are interested in estimators with a good overall performance the outlier size ω is sampled from a normal distribution with mean 0 and variance $a^2 \cdot \gamma(0),\ a \in \{5, 10, 20\}$. This produces some small perturbations, inconspicuous outliers which are favorable for monotone estimators, as well as very large outliers which redescenders can cope well with. Note that the outlier variance is proportional to the process variance. In general smaller values of a favor monotone estimators, while larger values favor redescenders and GK-approaches.

Furthermore, we contaminate an increasing number $n_0 \in \{5, 10, 15, 20, 25\}$ of values of the original time series of length $n = 100$ to investigate how many outliers an estimator can deal with. It was argued by Ma and Genton (2000) that estimators of the autocovariance cannot be expected to cope with more than 25% contaminated observations since one outlier can enter two pairs of observations entering the calculation of the correlation. Accordingly it is not reasonable to choose n_0 larger than 25. Moreover, we consider both isolated and patchy outliers, since these will have different effects. All results are based on 1000 simulation runs each.

We first treat the situation of isolated outliers, which drive the sample acf towards zero. The positions of the outliers are chosen at random for each time series. We first show the results for the AR[0.8] model with $a = 5$, which corresponds to an outlier variance of $25\gamma(0)$. As one can see in Figure 2.5, the empirical acf becomes useless already for $n_0 = 5$ outliers. In the same situation, some robust alternatives lose more than half of their efficiency. Nevertheless they are all preferable to the empirical acf. Their efficiency is at least 4.2 times larger.

Estimators which cope especially well with additive outliers are those based on robust filtering and to some extent also the GK approaches.

If one increases the outlier variance to $400\gamma(0)$, monotone estimators lose little more efficiency and get a somewhat larger bias, as expected, see Figure 2.6. Furthermore the filter methods and GK approaches gain efficiency compared to smaller outlier variances. For the latter this seems at first to be surprising since GK estimators are not known to be redescending, but it is consistent with the multivariate setting. The influence function derived in Ma and Genton (2001) tends to zero along the axes at least in the elliptical model. This means that a small fraction of outliers in only one dimension will have nearly no influence on the estimation as long as the outliers are large. In the case of isolated outliers we expect only one variable of each pair of observations to be contaminated and therefore a redescending behavior of the GK methods.

The estimators generally behave better for models with small absolute autocorrelations. This is not surprising, since the bias effect is more limited there. In models with rather small absolute autocorrelations other robust estimators like RA and GK approaches outperform the Filter-acf, which seems to behave especially well if the autocorrelations have large absolute values.

Patchy (consecutive) outliers increase the sample acf at small time lags towards one. For our simulations we add the same value ω generated from a $N(0, a^2 \cdot \gamma(0))$ distributed random variable to successive observations at times $\{51, \ldots, 50+n_0\}$. This resembles a temporarily level shift of the same height. We first look at the AR[0]-model with an outlier variance $\gamma(0)100$. Again the estimation by the empirical acf is useless already for $n_0 = 5$ outliers, see Figure 2.7. Robust estimators can cope much better with this situation and rarely preserve less than half of their efficiency, reaching values between 3.5 and 22 times the efficiency of the empirical acf in our experiments. Estimators based on SD and MCD and to some extent also the multivariate S-estimator perform well. Even for large amounts of outliers they are little biased and they lose only little efficiency. Different kinds of reweighting which boost efficiency in the clean model do not significantly increase the bias or vulnerability to outliers and should be preferred. Recall that GK approaches are not redescending under patchy outliers as can be seen in Figure 2.4.

If the acf attains large positive values, a few consecutive outliers can even improve the estimation by cancelling the small sample bias. All estimators behave better compared to the AR[0] model in these cases, with rank and RA-estimators improving most. Patchy outliers seem to have the largest impact if the acf contains values close to -1. We observe the largest maximal absolute biases and the smallest minimal efficiencies for the AR[-0.4] and AR[-0.8] models. The only exceptions are the filter methods which perform even better than in the AR[0] case. This agrees with higher efficiencies for larger absolute correlations in the absence of contamination.

2.3.3 Non-linear models

We also consider nonlinear models, specifically GARCH models, introduced by Bollerslev (1986). They allow the variance of the process to depend on the past observations and are popular for modelling financial time series (see for example Bollerslev et al., 1992; Duan, 1995). A GARCH process $(X_t)_{t\in\mathbb{Z}}$ of order $(1,1)$ is defined as

$$X_t = \sigma_t a_t \text{ with } \sigma_t^2 = \sigma_0^2 + \alpha_1 X_{t-1}^2 + \sum_{j=1}^{l} \beta_1 \sigma_{t-1}^2,$$

with parameters $\sigma_0, \alpha_1, \beta_1 \in \mathbb{R}_+$ and innovations $(a_t)_{t\in\mathbb{Z}}$ usually forming Gaussian or t-distributed white noise. In our simulations we consider a process of order (1,1) with $\sigma_0 = 0.05$, $\alpha_1 = 0.1$, $\beta_1 = 0.85$ and standard normal innovations $(a_t)_{t\in\mathbb{Z}}$ (abbreviated as GARCH[0.05,0.1,0.85]) which is a realistic parameter setting (see for example Lamoureux and Lastrapes, 1990). Results under patchy outliers can be seen in Figure 2.8. In the clean model robust procedures are a little more efficient than in the linear case, which might be due to the heavy tails of the marginal distribution of GARCH processes (Basrak et al., 2002). Under patchy and also under isolated outliers the results (not shown here) are comparable to the results for linear processes.

For modelling GARCH processes the autocorrelations of the squared process $(X_t^2)_{t\in\mathbb{Z}}$ are also of interest (see Baillie and Chung, 2001). Since estimation of the acf of $(X_t^2)_{t\in\mathbb{Z}}$ is harder, we choose a time series of length $n = 1000$. It turns out that most estimators are substantially biased, which might be a result of the skewed distribution of the process $(X_t^2)_{t\in\mathbb{Z}}$. Using parameters $\alpha_1 \in \{0, 0.1, 0.2\}$ and $\beta_1 \in \{0, 0.5, 0.7\}$ we find that in addition to the empirical acf only the RA approach with the Huber function and the median correlation yield an acceptable bias.

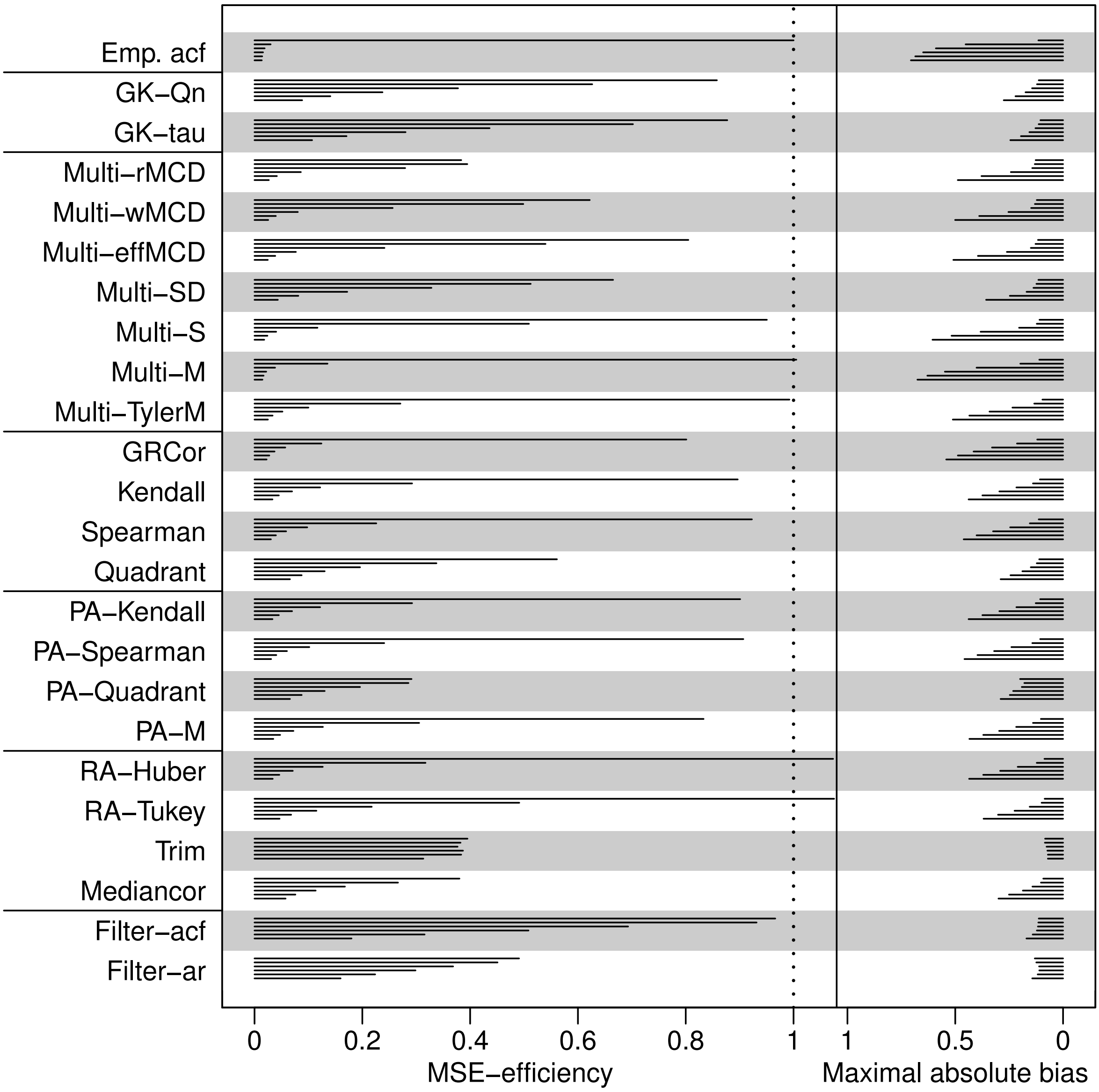

Figure 2.5: Efficiency (left) and bias (right) for a contaminated AR[0.8] model with $n = 100$ and $n_0 = 0, 5, 10, 15, 20, 25$ (from top to bottom in each panel) isolated outliers and $a = 5$.

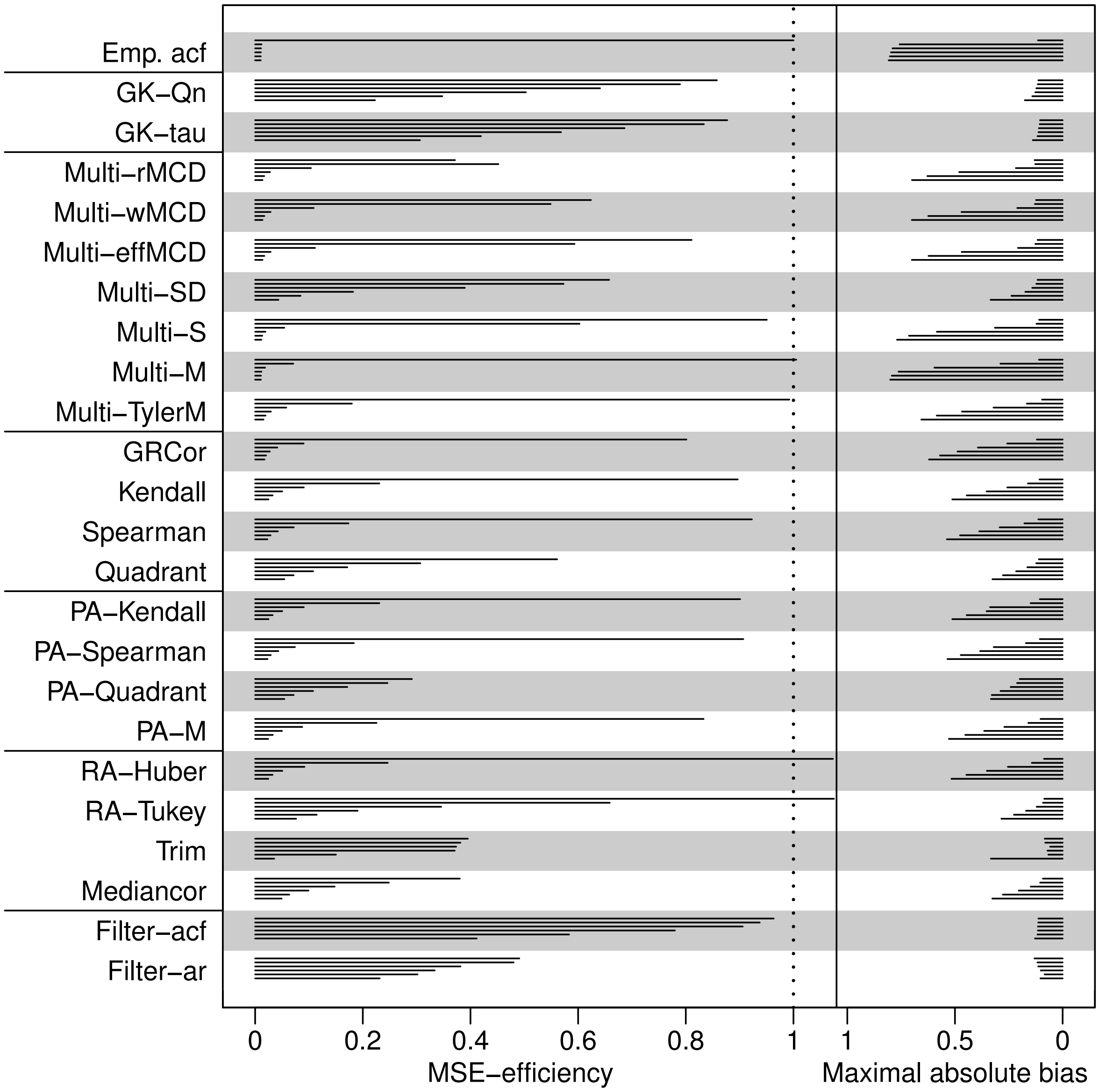

Figure 2.6: Efficiency (left) and bias (right) for a contaminated AR[0.8] model with $n = 100$ and $n_0 = 0, 5, 10, 15, 20, 25$ (from top to bottom in each panel) isolated outliers and $a = 20$.

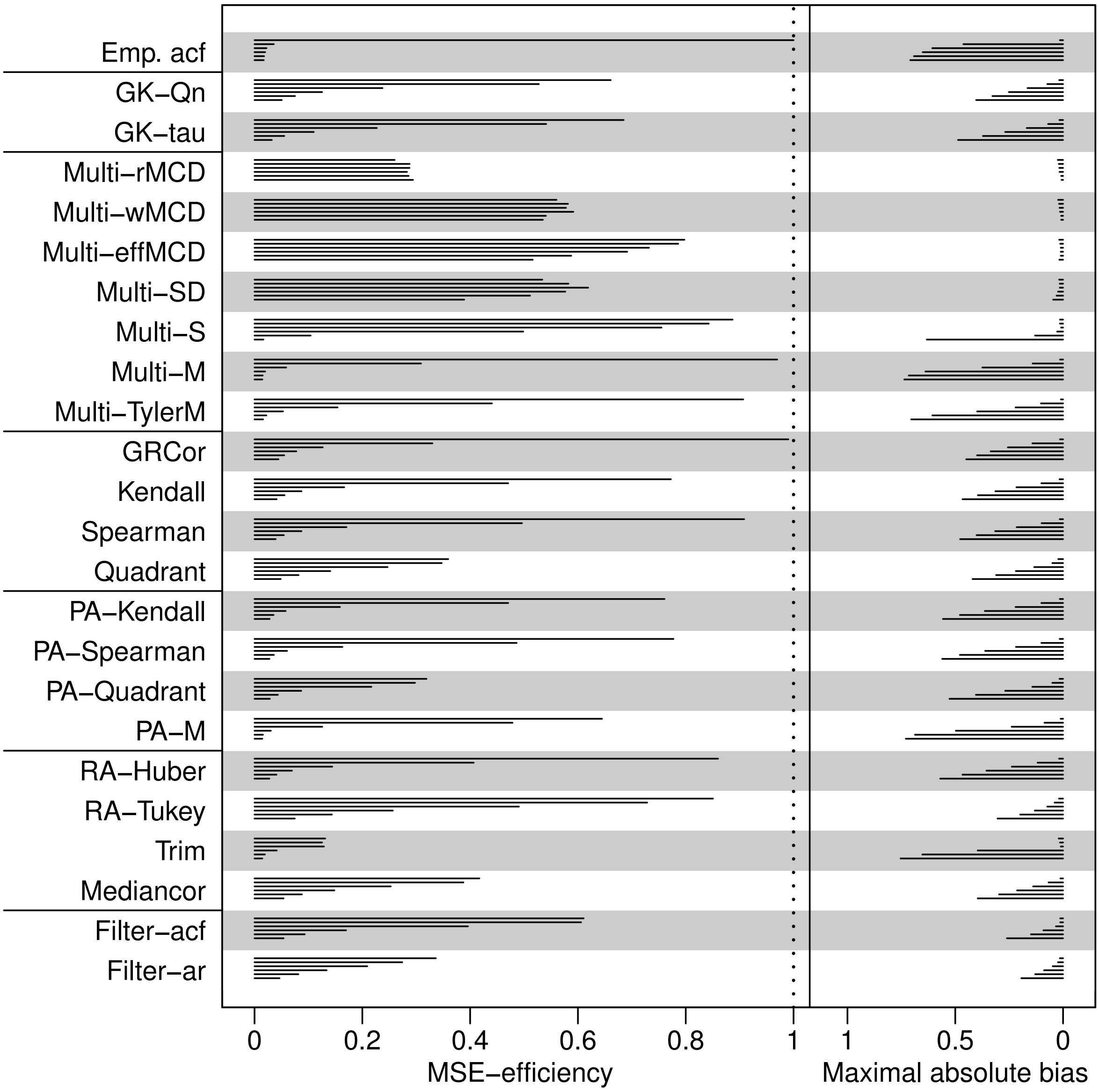

Figure 2.7: Efficiency (left) and bias (right) for a contaminated AR[0] model with $n = 100$ and outlier patches of length $n_0 = 0, 5, 10, 15, 20, 25$ and $a = 10$.

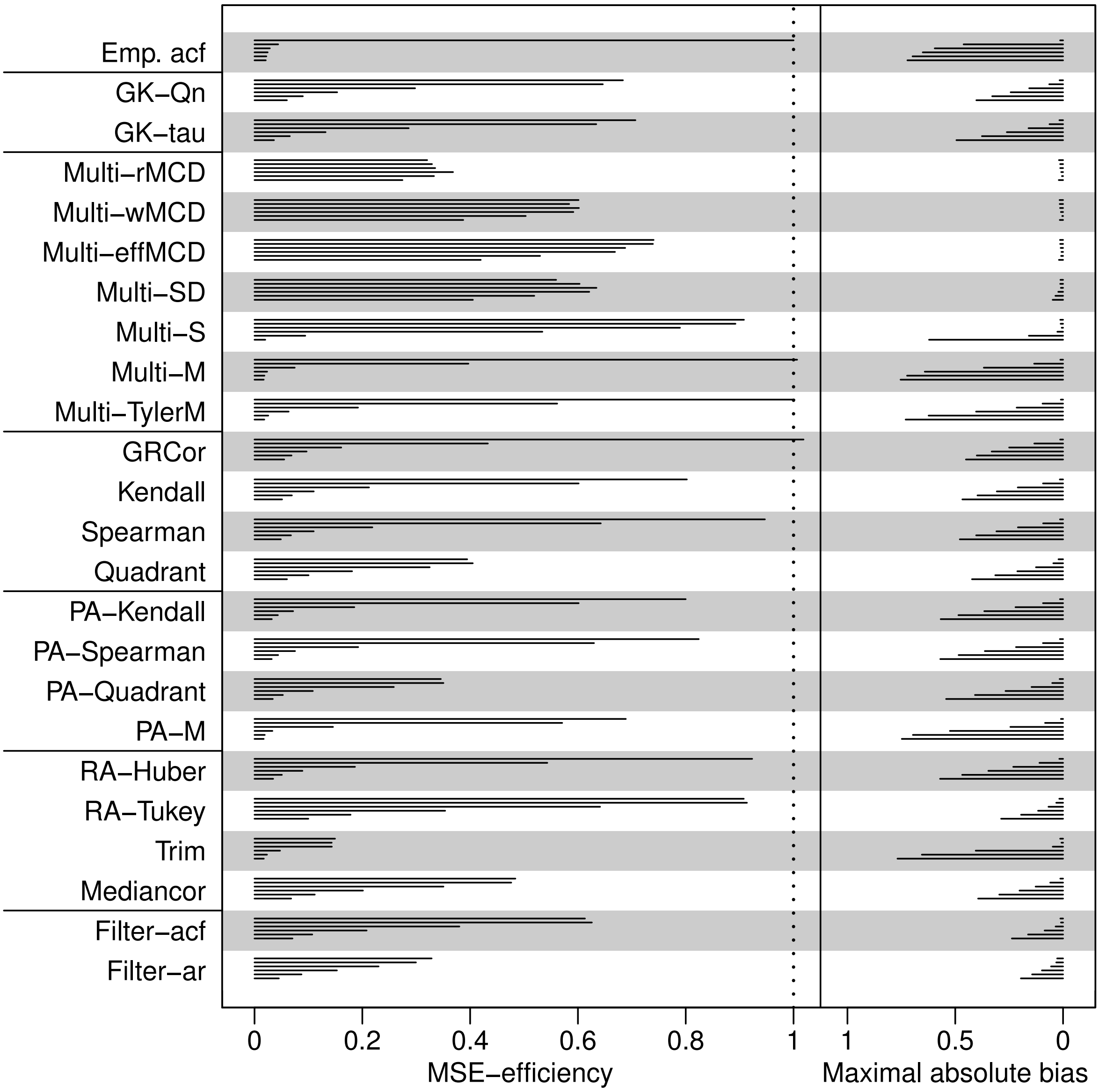

Figure 2.8: Efficiency under an outlier patch of length $n_0 = 0, 5, 10, 15, 20, 25$ (from top to bottom in each panel) under a GARCH[0.05,0.85,0.1] model with $n = 1000$ and $a = 10$.

2.3.4 Positive-semidefiniteness

We have mentioned the problem of positive-semidefiniteness repeatedly. Our simulations reveal that this is mainly a problem of little efficient estimators like quadrant correlation and the 50% trimming (median) approach. We never noticed problems for multivariate approaches except for the raw MCD, which occasionally produces indefinite estimations if the model is close to being non-stationary. It turns out that consistency corrections for the approaches based on univariate transformations often destroy definiteness. Whereas the difference between the original and the enforced positive-semidefinite estimation is negligible for the RA-estimators, we observed changes up to 0.08 for trimmed and median based correlation. There can be even greater discrepancies for the Filter-AR estimator, which might be caused by some instability of our implementation of this procedure. We rarely noticed indefinite estimations by the variance based approaches. Enforcing positive-semidefiniteness increases the efficiency of trimmed estimators slightly.

2.4 Conclusion

Many of the proposals for robust autocorrelation estimation are borrowed from the usual correlation estimation applied to all pairs of observations (X_t, X_{t+h}) at a certain time lag h, with the intention of carrying over good robustness properties and high efficiency under normality to the time series context. A problem arising there is that every outlier can enter two pairs of observations, so that the number of contaminated pairs can be up to twice the number of outliers. This problem does not arise for estimators which filter the time series before the acf estimation. However, these do not respect the serial dependence structure like the RA-estimators or they are computationally heavy like the robust filter algorithm.

There is a great interest in robust time series analysis nowadays. There is also a new proposal by Chang and Politis (2014) on autocorrelation estimation based on the idea of estimating $\rho(h)$ by regressing X_{t+h} robustly on X_t. Extra manipulations are needed to guarantee that such estimates are positive-semidefinite and do not exceed 1. Moreover, there are many candidate robust regression techniques available, so that a careful inspection of this proposal would have been beyond the scope of this paper.

Our simulation study confirms that even a small fraction of contamination can make the empirical acf useless. The robust filter algorithms yield good results even in case of many isolated or patchy outliers, but have a lower efficiency if there is little serial correlation. Estimation based on a reweighted MCD is favorable, if there are patchy outliers. The approach based on the Stahel-Donoho estimator means a good compromise, but it is computationally demanding. If one looks for a relatively quick estimator, the approach based on robust variances seems to be a good choice, since they also generally yield good results. A possible lack of positive-semidefiniteness can easily be fixed by a projection algorithm.

Our simulation results are based on additive outliers of random size ω and therefore represent a kind of overall performance for different outlier sizes. In simulations not reported here we also consider other outlier scenarios with fixed outlier sizes. The worst case biases and efficiencies are generally worse there than those presented here. Nevertheless, the results are qualitatively rather similar.

It needs to be kept in mind that in the simulations reported here we focus on the case of innovations from a contaminated Gaussian or at least continuous-symmetric distribution. Results look different e.g. for count time series as reported in Fried et al. (2014), where rank based estimators performed rather well.

Chapter 3

Spatial sign correlation

3.1 Introduction

In this chapter we present a new estimator for the correlation coefficient and derive its asymptotic properties. The new proposal is based on the spatial sign covariance matrix. The *spatial sign* of a multivariate observation is its projection (after a suitable centering) onto the p-dimensional unit sphere. Spatial signs play an important role in robust multivariate data analysis. Since every observation is basically shrunk to length 1, the impact of any contamination is bounded. Spatial signs have been used, e.g., for robust tests of multivariate location (e.g. Möttönen et al., 1997), tests of independence (Taskinen et al., 2003), testing for sphericity (Sirkiä et al., 2009) or canonical correlation analysis (Taskinen et al., 2006). Using spatial signs as score function in estimation leads to the spatial median as a multivariate location estimator or, in the regression setting, to the least absolute deviation (LAD) regression. For a recent overview of spatial sign methods see Oja (2010).

The *spatial sign covariance matrix* (SSCM) is simply the covariance matrix of the spatial signs of the (suitably centered) observations. It is known that, within symmetric data models, the SSCM consistently estimates the eigenvectors of the covariance matrix, but not the eigenvalues (see for example Marden, 1999). The connection between the eigenvalues of the population SSCM and the covariance matrix is explicitly known in the special case of two-dimensional elliptical distribution, see Vogel et al. (2008) and Croux et al. (2010). We use this relationship to robustly estimate a two-dimensional covariance matrix (up to scale) based on the SSCM and hence devise a correlation estimator, which we call *spatial sign correlation.* We further derive the asymptotic distributions and influence functions of the SSCM and the spatial sign correlation. It turns out that the asymptotic variance of the spatial sign correlation can get arbitrarily large if the ratio of the marginal scales get arbitrarily large. We therefore propose a two stage procedure which first standardizes the data marginally and then computes the spatial sign correlation. The asymptotic variance of the two stage spatial sign correlation equals then that of the ordinary spatial sign correlation for the most favourable case of equal marginal scales.

Finally we introduce two generalizations to estimate the multivariate correlation matrix. The first one fills the off-diagonal positions of the matrix estimate with the bivariate spatial sign correlation coefficients of all pairs of variables. The second one which we call the *positive definite spatial sign correlation matrix* uses the relationship between the eigenvalues of the population SSCM and the covariance matrix under elliptical distributions of arbitrary dimensions. Though an explicit formula mapping the eigenvalues onto each other does not seem to exist except in the bivariate case, the eigenvalues of the SSCM can be expressed as one-dimensional integrals given the eigenvalues of the covariance matrix. This representation can be inverted by a fix point algorithm and therefore used to estimate the correlation matrix. Simulations and theoretical results suggest that the positive definite spatial sign correlation matrix is more efficient but also more sensitive against outliers than the bivariate approach.

This chapter is structured as follows. We gather results concerning the SSCM including representations of the eigenvalues in Section 3.2. We use them to derive the spatial sign correlation coefficient in Section 3.3 and calculate its asymptotic distribution as well as its influence function and gross error sensitivity. Section 3.4 is devoted to the two stage spatial sign correlation. Besides its asymptotic distribution we state a variance stabilising transformation in the same manner as the Fisher z-transform, which enables us to construct very accurate confidence intervals even for small sample sizes. In Section 3.5 we compare the spatial sign correlation analytically and in simulations with other robust correlation estimators which are commonly used. Section 3.6 is concerned with the positive definite spatial sign correlation. It contains a description of the fixed point algorithm, simulations regarding the efficiency of the estimator and its influence function under special assumptions, which indicates the sensitivity of the positive definite spatial sign correlation matrix in high dimensions. All proofs are deferred to Section 3.8.

We close this section by introducing some recurrent terms and notation. By vec we denote the operator, which stacks the columns of a matrix from left to right underneath each other, and by $\otimes$ the Kronecker product (e.g. Magnus and Neudecker, 1999, Sec. 2). Both are connected by the identity $\operatorname{vec}(ABC) = (C^T \otimes A)\operatorname{vec} B$. Furthermore we denote by $\mathbb{X}_n = (\boldsymbol{X}_1, ..., \boldsymbol{X}_n)^T$ the $n \times p$ data matrix containing the p-dimensional observations $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ as rows. In order to study the properties of the new estimator analytically we will assume the data to stem from the elliptical model. A continuous distribution F on $\mathbb{R}^p$ is said to be *elliptical* if it has a Lebesgue-density f of the form

$$f(\boldsymbol{x}) = \det(V)^{-1/2} g\{(\boldsymbol{x} - \boldsymbol{\mu})^T V^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\} \tag{3.1}$$

for some $\boldsymbol{\mu} \in \mathbb{R}^p$ and symmetric, positive definite $p \times p$ matrix V. We call $\boldsymbol{\mu}$ the *location* or *symmetry center*, V the *shape matrix*, since it describes the shape of the elliptical contour lines of the density, and the function $g : [0, \infty) \to [0, \infty)$ the *elliptical generator* of F. The class of all continuous elliptical distributions F on $\mathbb{R}^p$ having location $\boldsymbol{\mu}$ and shape V is denoted by $\mathscr{E}_p(\boldsymbol{\mu}, V)$. The shape matrix V is unique only up to scale, that is, $\mathscr{E}_p(\boldsymbol{\mu}, V) = \mathscr{E}_p(\boldsymbol{\mu}, cV)$ for any $c > 0$. For scale-free functions of V, such as correlations, which we consider here, this ambiguity is irrelevant. A common view on the *shape* of an elliptical distribution is to treat it as an equivalence class of positive definite random matrices being proportional to each other. We adopt this notion here: in the results of this exposition, V can be any representative from its equivalence class. If second moments exist, one can always take the covariance matrix, or any suitably scaled multiple of it. However, the results are more general, the existence of second, or even first, moments is not required. Throughout the paper we let

$$V = U \Lambda U^T \tag{3.2}$$

denote an eigenvalue decomposition of V, where U is an orthogonal matrix containing the eigenvectors of V as columns and $\Lambda = \operatorname{diag}(\lambda_1, ..., \lambda_p)$ is such that $0 < \lambda_p \leq \ldots \leq \lambda_1$. It furthermore holds that for $\boldsymbol{X} \sim F \in \mathscr{E}_p(\boldsymbol{\mu}, V)$, there exists a spherical random variable $\boldsymbol{Y}$ such that

$$\boldsymbol{X} = U \Lambda^{\frac{1}{2}} \boldsymbol{Y} + \boldsymbol{\mu}, \tag{3.3}$$

with U and Λ as in (3.2). We use $||\cdot||$ to denote the L_2 norm of a vector.

3.2 The spatial sign covariance matrix

We define the spatial sign covariance matrix of a multivariate distribution and derive its connection to the shape matrix V in case of elliptical distributions. For $\boldsymbol{x} \in \mathbb{R}^p$ define the *spatial sign*

$\boldsymbol{s}(\boldsymbol{x})$ of $\boldsymbol{x}$ as $\boldsymbol{s}(\boldsymbol{x}) = \boldsymbol{x}/||\boldsymbol{x}||$ if $\boldsymbol{x} \neq \mathbf{0}$ and $\boldsymbol{s}(\boldsymbol{x}) = \mathbf{0}$ otherwise. Let $\boldsymbol{X}$ be a p-dimensional random vector ($p \geq 2$) having distribution F. We call $\boldsymbol{\mu}(F) = \boldsymbol{\mu}(\boldsymbol{X}) = \arg\min_{\boldsymbol{\mu}\in\mathbb{R}^p} \mathbb{E}(||\boldsymbol{X} - \boldsymbol{\mu}|| - ||\boldsymbol{X}||)$ the *spatial median* and, following the terminology of Visuri et al. (2000),

$$S(F) = S(\boldsymbol{X}) = \mathbb{E}\left(\boldsymbol{s}(\boldsymbol{X} - \boldsymbol{\mu}(F))\boldsymbol{s}(\boldsymbol{X} - \boldsymbol{\mu}(F))^T\right) \tag{3.4}$$

the *spatial sign covariance matrix (SSCM)* of F (or $\boldsymbol{X}$). If there is no unique minimizing point of $\mathbb{E}(||\boldsymbol{X} - \boldsymbol{\mu}|| - ||\boldsymbol{X}||)$, then define $\boldsymbol{\mu}(F)$ as the barycenter of the minimizing set. This may only happen if F is concentrated on a line. For results on existence and uniqueness of the spatial median see Haldane (1948), Kemperman (1987), Milasevic and Ducharme (1987) or Koltchinskii and Dudley (2000). If the first moments of F are finite, then the spatial median allows the more descriptive characterization as $\arg\min_{\boldsymbol{\mu}\in\mathbb{R}^p} \mathbb{E}||\boldsymbol{X} - \boldsymbol{\mu}||$. Let $\mathbb{X}_n = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^T$ be a data sample of size n, where the $\boldsymbol{X}_i$, $i = 1, ..., n$, are i.i.d., each with distribution F. Define

$$\hat{S}_n(\mathbb{X}_n; \boldsymbol{t}) = \underset{i=1,\ldots,n}{\text{ave}}\ \boldsymbol{s}(\boldsymbol{X}_i - \boldsymbol{t})\boldsymbol{s}(\boldsymbol{X}_i - \boldsymbol{t})^T \tag{3.5}$$

where $\boldsymbol{t} \in \mathbb{R}^p$. Choosing $\boldsymbol{t} = \boldsymbol{\mu}(F)$, we call the estimator $\hat{S}_n(\mathbb{X}_n; \boldsymbol{\mu}(F))$ the *empirical SSCM with known location*. However, the location is usually unknown, and $\boldsymbol{t}$ has to be replaced by a suitable location estimator $(\boldsymbol{t}_n)_{n\in\mathbb{N}}$, and we refer to $\hat{S}_n(\mathbb{X}_n; \boldsymbol{t}_n)$ as the *empirical SSCM with unknown location*. The canonical location functional in this case is the *(empirical) spatial median*

$$\hat{\boldsymbol{\mu}}_n = \hat{\boldsymbol{\mu}}_n(\mathbb{X}_n) = \min_{\boldsymbol{\mu}\in\mathbb{R}^p} \sum_{i=1}^{n} ||\boldsymbol{X}_i - \boldsymbol{\mu}||.$$

Under regularity conditions (the data points do not lie on a line and none of them coincides with $\hat{\boldsymbol{\mu}}_n$, see Kemperman (1987), p. 228) the spatial signs with respect to the empirical spatial median are centered, i.e. $\text{ave}_{i=1}^n\, \boldsymbol{s}(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_n) = \mathbf{0}$. Hence, the empirical spatial sign covariance matrix $\hat{S}_n(\mathbb{X}_n; \hat{\boldsymbol{\mu}}_n)$ is indeed the covariance matrix of the spatial signs if the latter are taken with respect to the spatial median.

Within the elliptical model there is, up to scale, a one-to-one connection between $S(F)$ and the parameter V: both share the same eigenvectors and the ordering of the respective eigenvalues: $S(F) = U\Delta U^T$, where $\Delta = \text{diag}(\delta_1, ..., \delta_p)$ is a diagonal matrix with $0 < \delta_p \leq \ldots \leq \delta_1$. This makes the spatial sign covariance matrix particularly popular for robust principal component analysis (e.g. Marden, 1999; Locantore et al., 1999; Croux et al., 2002; Gervini, 2008). Other applications are direction-of-arrival estimation (Visuri et al., 2001), or testing sphericity in the elliptical model (Sirkiä et al., 2009). The map between the eigenvalues of V and $S(F)$ is only known explicitly for $p = 2$, where

$$\delta_j = \sqrt{\lambda_j}/(\sqrt{\lambda_1} + \sqrt{\lambda_2}), \quad j = 1, 2. \tag{3.6}$$

This result first appears in a similar form in Visuri et al. (2000) and has also been used by Vogel et al. (2008) and Croux et al. (2010), but neither of these articles provide a proof, which can finally be found in Dürre et al. (2015b). In case of general p decomposition (3.3) of an elliptic random variable enables the following representation of the eigenvalues of $S(F)$

$$\delta_i = E\left\{\lambda_i Y_i^2 \left(\textstyle\sum_{j=1}^p \lambda_j Y_j^2\right)^{-1}\right\} \tag{3.7}$$

where $Y = (Y_1, \ldots, Y_p)$ has a spherical distribution. However, such p-dimensional integrals are hard to approximate numerically. The next proposition offers a much simpler description.

Proposition 1. *(Dürre et al., 2016) Let* $F \in \mathscr{E}_p(\boldsymbol{\mu}, V)$ *and* $V = U\Lambda U^T$ *denote an eigenvalue decomposition of* V *with* $0 \leq \lambda_p \leq \ldots \leq \lambda_1$*. Then the eigenvalues* $0 \leq \delta_p \leq \ldots \leq \delta_1$ *of* $S(F)$ *have the following representation:*

$$\delta_j = \frac{\lambda_j}{2} \int_0^\infty \frac{1}{(1+\lambda_j x)\prod_{i=1}^p (1+\lambda_i x)^{1/2}} dx, \qquad 1 \leq j \leq p. \tag{3.8}$$

The p-dimensional integral (3.7) is unfeasible for numerical approximations. Solving the one-dimensional integral given in Proposition 1 numerically, the population SSCM can be computed in any dimension. Using the function integrate() in R (R Core Team, 2016), we found it to work without problems for $p = 10{,}000$.

We use formula (3.8), which is implemented in R in the package sscor (Dürre and Vogel, 2016b), to get an impression how the eigenvalues of $S(\mathbf{X})$ look like in comparison to those of V. As mentioned before V is only uniquely defined up to scale. For a better comparability we look at the eigenvalues of the trace standardized shape matrix $V_1 = V/tr(V)$ (keep in mind that the trace of $S(F)$ equals 1 per definition). We first look at equidistantly spaced eigenvalues

$$\lambda_i = \frac{2(p+1-i)}{p(p+1)}, \quad i = 1, \ldots, p,$$

for different $p = 3,\ 11,\ 101$.

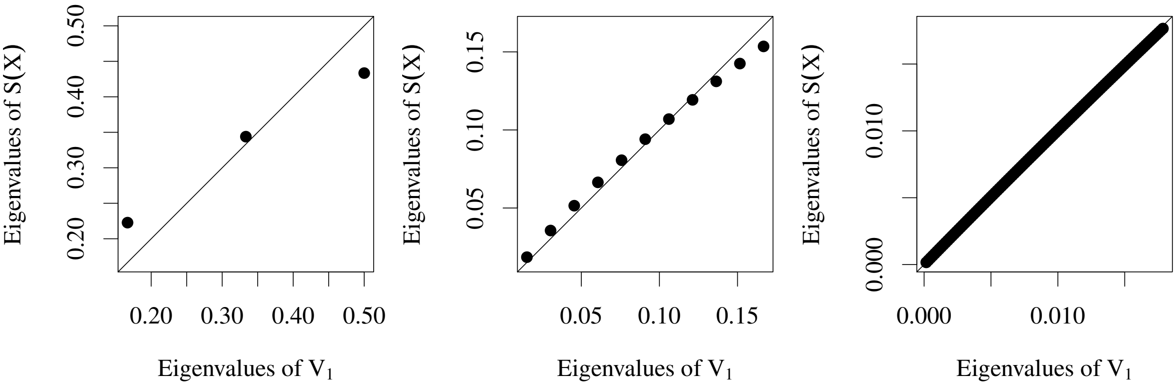

Figure 3.1: Eigenvalues of the SSCM w.r.t. the corresponding eigenvalues of the shape matrix in the equidistant setting $p = 3$ (left), $p = 11$ (centre) and $p = 101$ (right).

The magnitude of the eigenvalues necessarily decreases as p increases, since $\sum_{i=1}^p \lambda_i = \sum_{i=1}^p \delta_i = 1$ per definition of V_1 and $S(\mathbf{X})$. As one can see in Figure 3.1, the eigenvalues of $S(\mathbf{X})$ and V_1 approach each other for increasing p. In fact the maximal absolute difference for $p = 101$ is roughly $2 \cdot 10^{-4}$. In the second scenario, we take $p-1$ equidistantly spaced eigenvalues and one eigenvalue 5 times larger than the rest, i.e.,

$$\lambda_i = \begin{cases} \frac{5(p-1)}{p((p+1)/2+5)-5} & i = 1, \\ \frac{p-i}{p((p+1)/2+5)-5} & i = 1, \ldots, p-1. \end{cases}$$

This models the case where the dependence is mainly driven by one principal component.

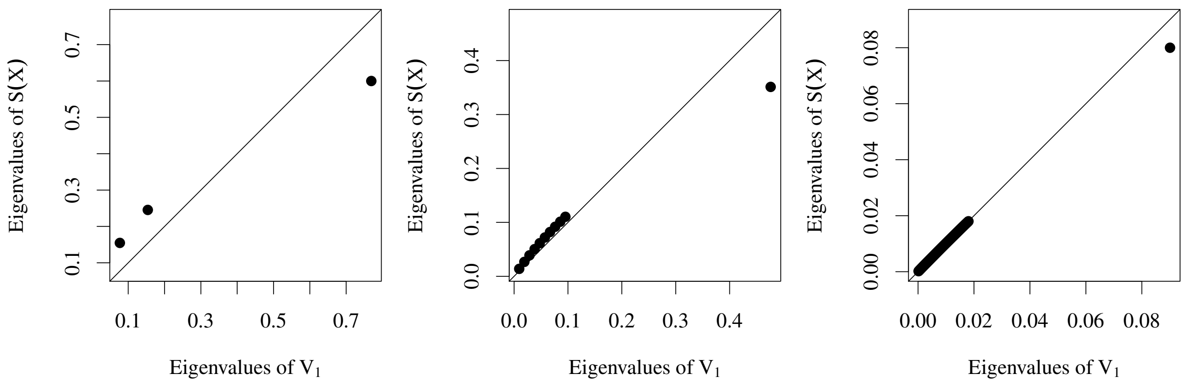

Figure 3.2: Eigenvalues of the SSCM wrt the corresponding eigenvalues of shape matrix in the setting of one large eigenvalue for $p = 3$ (left), $p = 11$ (centre) and $p = 101$ (right).

As one can see in Figure 3.2, the distance between the two largest eigenvalues is smaller for $S(\mathbf{X})$ than for V_1. This is not surprising, since it is proven in Proposition 2 of Dürre et al. (2016) that

$$\delta_i/\delta_j \leq \lambda_i/\lambda_j \text{ for } 1 \leq i < j \leq p \text{ and } \lambda_j > 0.$$

Thus in general, the eigenvalues of the SSCM are less separated than those of V_1, which is one reason why the use of the SSCM for robust principal component analysis has been questioned (e.g. Bali et al., 2011; Magyar and Tyler, 2014). However, the differences appear to be generally small in higher dimensions.

The next paragraph concerns the asymptotic behaviour of the empirical SSCM. Let therefore $\boldsymbol{X}_1, \ldots \boldsymbol{X}_n$ be i.i.d. with distribution F, which is not necessarily elliptical. Under the assumptions that $\boldsymbol{t}_n$ is strongly consistent for $\boldsymbol{\mu}$ and $\mathbb{E}\left(||\boldsymbol{X}_1 - \boldsymbol{\mu}||^{-1}\right) < \infty$ Dürre et al. (2014) showed that then also the spatial sign covariance matrix is strongly consistent

$$\hat{S}_n(\mathbb{X}_n; \boldsymbol{t}_n) \xrightarrow{a.s.} S(F).$$

While the condition concerning the location estimator $\boldsymbol{t}_n$ is quite usual, reflects the moment condition the specific construction of the SSCM, more precisely the discontinuity of the spatial sign at the origin. However the condition is very mild and for example fulfilled if the density f is bounded.

If one assumes additionally that $\sqrt{n}(\boldsymbol{t}_n - \boldsymbol{\mu})$ converges in distribution, $\mathbb{E}\left(||\boldsymbol{X}_1 - \boldsymbol{\mu}||^{-3/2}\right) < \infty$ and $(\boldsymbol{X}_1 - \boldsymbol{\mu}) \overset{\mathscr{L}}{=} -(\boldsymbol{X}_1 - \boldsymbol{\mu})$ then the SSCM is asymptotically normal, so there exists a symmetric and non-negative definite symmetric matrix $W_S \in \mathbb{R}^{p \times p}$ such that

$$\sqrt{n}\,\text{vec}\left\{\hat{S}_n(\mathbb{X}_n; \boldsymbol{t}_n) - S(F)\right\} \xrightarrow{d} N_{p^2}\left(\mathbf{0}, W_S\right),$$

see Dürre et al. (2014), where also alternative assumptions to guarantee strong consistency and asymptotic normality are given. The symmetry condition $(\boldsymbol{X}_1 - \boldsymbol{\mu}) \overset{\mathscr{L}}{=} -(\boldsymbol{X}_1 - \boldsymbol{\mu})$ is not necessary, but it ensures that the location estimation does not enter into the asymptotics of the SSCM, in particular that it does not influence the form of W_S. It is fulfilled for example for elliptical distributions. In this case one can specify W_S by

$$W_S = (U \otimes U)\left\{\Gamma - \text{vec}\,\Delta(\text{vec}\,\Delta)^\top\right\}(U \otimes U)^\top,$$

with

$$\Gamma = E\left\{\mathrm{vec}\left(\frac{\Lambda^{1/2}YY^\top\Lambda^{1/2}}{Y^\top\Lambda Y}\right)\mathrm{vec}\left(\frac{\Lambda^{1/2}YY^\top\Lambda^{1/2}}{Y^\top\Lambda Y}\right)^\top\right\}.$$

Due to the spherical symmetry of Y, $p(p^3-3p+2)$ of the p^4 matrix entries of Γ are zero. The remaining $p(3p-2)$ entries consist of at most $(p+1)p/2$ distinct values, with the upper bound being achieved if the eigenvalues $\lambda_1,\ldots,\lambda_p$ of V_0 are mutually distinct. Letting

$$\eta_{ij} = E\left\{\lambda_i Y_i^2 \lambda_j Y_j^2 \left(\textstyle\sum_{j=1}^p \lambda_j Y_j^2\right)^{-2}\right\}, \qquad 1 \le i,j \le p, \tag{3.9}$$

we have for $1 \le i < j \le p$, each η_{ij} appears six times in Γ, that is at the positions $\{(i-1)p+j,\, (i-1)p+j)\}$, $\{(i-1)p+i,\, (j-1)p+j)\}$, $\{(i-1)p+j,\, (j-1)p+i)\}$, and the same with i and j interchanged. For $1 \le i \le p$, each η_{ii} appears once at position $\{(i-1)p+i,\, (i-1)p+i)\}$. In the two-dimensional case Dürre et al. (2015b) calculate the expectations (3.9) explicitly by applying the residue theorem and deduce

$$W_S = \frac{-\lambda_1\lambda_2 + \frac{1}{2}\sqrt{\lambda_1\lambda_2}(\lambda_1+\lambda_2)}{(\lambda_1-\lambda_2)^2}(U\otimes U)W_0(U\otimes U)^T \tag{3.10}$$

with

$$W_0 = \begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix},$$

if $\lambda_1 \ne \lambda_2$ respectively $W_S = \frac{1}{8}W_0$ for $\lambda_1 = \lambda_2$. The next proposition characterizes the integrals (3.9) for general p.

Proposition 2. *(Dürre et al., 2016) Let $\boldsymbol{Y} = (Y_1,\ldots,Y_p)$ be spherical distributed and $\lambda_1 \ge \ldots \ge \lambda_p \ge 0$, then $\eta_{i,j}$ defined in (3.9) possesses the following representation:*

$$\eta_{ij} = \begin{cases} \frac{\lambda_i\lambda_j}{4}\int_0^\infty \frac{x}{(1+\lambda_i x)(1+\lambda_j x)\prod_{k=1}^p(1+\lambda_k x)^{1/2}}dx, & i \ne j \\ \frac{3\lambda_i^2}{4}\int_0^\infty \frac{x}{(1+\lambda_i x)^2\prod_{k=1}^p(1+\lambda_k x)^{1/2}}dx, & i = j. \end{cases}$$

In the case of two distinct eigenvalues, Magyar and Tyler (2014) investigate the asymptotic efficiency of the SSCM eigenspace projections by employing a representation of the eigenvalues δ_i and the η_{ij}-terms by means of the Gauss hyperbolic function. Proposition 2 allows to quantify the asymptotic efficiency of the SSCM and any analysis build upon it in the general setting.

3.3 A spatial sign based estimator for the correlation coefficient

In the following let $\boldsymbol{X}_i = (X_i, Y_i)^T$, $i = 1,\ldots,n$, be an i.i.d. sample from $F \in \mathscr{E}_2(\boldsymbol{\mu}, V)$. Denoting the entries of V by v_{ij}, we want to estimate the parameter

$$\rho = v_{12}/\sqrt{v_{11}v_{22}}.$$

We call ρ the *generalized correlation coefficient* of the elliptical distribution F, since it coincides with the correlation coefficient if second moments are finite. In a slight abuse of notation, we will refer to ρ simply as the correlation (coefficient) of F in the following. Equation (3.6) from the previous section gives rise to an estimator of ρ constructed as follows: compute the SSCM $\hat{S}_n = \hat{S}_n(\mathbb{X}_n; \hat{\boldsymbol{\mu}}_n)$, perform an eigenvalue decomposition $\hat{S}_n = \hat{U}_n\hat{\Delta}_n\hat{U}_n^T$ with $\hat{\Delta}_n = \mathrm{diag}(\hat{\delta}_1, \hat{\delta}_2)$ and compute the matrix $\hat{V}_n = \hat{U}_n\hat{\Lambda}_n\hat{U}_n^T$ with $\hat{\Lambda}_n = \mathrm{diag}(\hat{\lambda}_1, \hat{\lambda}_2)$ and $\hat{\lambda}_1 = \hat{\delta}_1/\hat{\delta}_2$, $\hat{\lambda}_2 = \hat{\delta}_2/\hat{\delta}_1$.[1]

[1] The overall scaling of $\hat{V}_n$ is, of course, irrelevant for the correlation, and its eigenvalues $\hat{\lambda}_1$ and $\hat{\lambda}_2$ may as well be chosen differently. By setting $\sqrt{\lambda_1} + \sqrt{\lambda_2} = 1$ one can see that the ratio has to satisfy $\hat{\lambda}_1/\hat{\lambda}_2 = (\hat{\delta}_1/\hat{\delta}_2)^2$.

Finally compute the correlation coefficient from the matrix $\hat{V}_n$, i.e. let $\hat{\rho}_n = \hat{v}_{12}/\sqrt{\hat{v}_{11}\hat{v}_{22}}$. In dimension two, the eigenvalue decomposition can be computed explicitly with justifiable effort which results in:

$$\hat{\rho}_n = \frac{\hat{s}_{12}}{\sqrt{(\hat{s}_{12}^2 + \hat{s}_{11}^2)(\hat{s}_{12}^2 + (1 - \hat{s}_{11})^2)}}, \tag{3.11}$$

where $\hat{s}_{ij}$ denote the entries of $\hat{S}_n$. We call $\hat{\rho}_n$ the *spatial sign correlation coefficient*. This must not be confused with the correlation of the spatial signs of the observations. This would be $\hat{\rho}_{\mathrm{SSCM}} = \hat{s}_{12}/\sqrt{\hat{s}_{11}\hat{s}_{22}}$. Also note that knowing $\hat{\rho}_{\mathrm{SSCM}}$ alone is not sufficient for computing $\hat{\rho}_n$. The next proposition deals with the asymptotic properties of $\hat{\rho}_n$.

Proposition 3. *(Dürre et al., 2015b) Let $F \in \mathscr{E}_2(\boldsymbol{\mu}, V)$ have a bounded density at $\boldsymbol{\mu}$. Then, as $n \to \infty$,*

(1) $\hat{\rho}_n \xrightarrow{a.s.} \rho$, and

(2) $\sqrt{n}(\hat{\rho}_n - \rho) \xrightarrow{d} N\left(0,\, (1-\rho^2)^2 + \frac{1}{2}\left(a + a^{-1}\right)(1-\rho^2)^{3/2}\right)$,
where $a = \sqrt{v_{11}/v_{22}}$ is the root of the ratio of the diagonal elements of V.

Proposition 3 (2) gives the asymptotic variance $ASV(\hat{\rho}_n)$ as a function of the true correlation ρ and the ratio of the diagonal elements of the shape matrix V. The elliptical generator g, cf. (3.1), does not enter, which may be phrased as "$\hat{\rho}_n$ is asymptotically distribution-free within the elliptical model". It is furthermore consistent and asymptotically normal without any moment condition.

In the next step, we examine the influence function of the spatial sign correlation. The influence function is based on the notion that estimators are statistical functionals working on distributions. The specific estimate computed from the data set $\mathbb{X}_n$ is then the functional applied to the corresponding empirical distribution. We use $\hat{S}$ and $\hat{\rho}$ to denote the statistical functionals corresponding to the SSCM and the spatial sign correlation, respectively.[2] The influence function $IF(\boldsymbol{x}, \hat{\rho}, F)$ describes the effect of an infinitesimal small contamination at point $\boldsymbol{x}$ on the functional $\hat{\rho}$ if the latter is evaluated at distribution F. It is an important tool describing the robustness properties of estimators. For a precise definition, interpretation and further details, see, e.g., Hampel et al. (1986) or Huber and Ronchetti (2009).

Proposition 4. *(Dürre et al., 2015b) Let $F \in \mathscr{E}_2(\boldsymbol{\mu}, V)$. Then $IF(\boldsymbol{x}, \hat{\rho}, F) =$*

$$\frac{-\left\{\left(a^2+1\right)\rho\sqrt{1-\rho^2}+2\,a\,\rho\left(1-\rho^2\right)\right\}\left(a^2\,y^2+x^2\right)-\left\{\left(a^4+6\,a^2+1\right)\left(\rho^2-1\right)+2\,a\left(a^2+1\right)\sqrt{1-\rho^2}\left(\rho^2-2\right)\right\}x\,y}{\left\{2\,a^2\sqrt{1-\rho^2}+a\left(a^2+1\right)\right\}\left(y^2+x^2\right)},$$

where $\boldsymbol{x} = (x, y)^T$ and a and ρ are as in Proposition 3.

The influence function for $a = 1$ and $\rho = 0$ is illustrated in Figure 3.3 on the right. It has a discontinuity at the origin and is bounded. Its extreme values ±2 are attained on the diagonals. Furthermore, $IF(\boldsymbol{x}, \hat{\rho}, F)$ is bounded in $\boldsymbol{x}$ for any fixed values a and ρ, but it may get arbitrarily large as a varies. A robustness index that is derived from the influence function is the gross-error sensitivity (GES), defined as $GES(\hat{\rho}, F) = \sup_{\boldsymbol{x} \in \mathbb{R}^2} |IF(\boldsymbol{x}, \hat{\rho}, F)|$. For $a = 1$, we obtain

$$GES(\hat{\rho}, F) = \frac{\left\{\left(\rho^2-1\right)\left(-\rho^4+8\,\rho^2+4\sqrt{1-\rho^2}\left(\rho^2-2\right)-8\right)\right\}^{1/2}+|\rho|\left(\sqrt{1-\rho^2}-\rho^2+1\right)}{\sqrt{1-\rho^2}+1}.$$

[2] We use the notational convention that functionals working on distributions wear hats, but do not carry subscripts. So for instance $S(F)$ and $\hat{S}(F)$ denote the same thing, but the different notation invokes slightly different views: $S(F)$ denotes a parameter of the distribution F, while $\hat{S}(F)$ is the functional $\hat{S}$ evaluated at F.

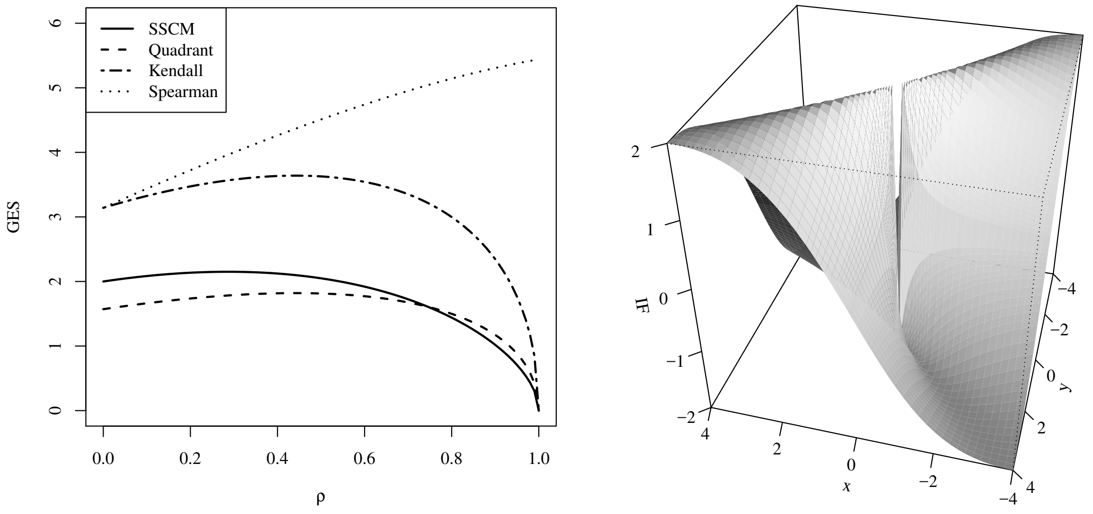

Figure 3.3: The gross error sensitivity (GES) of the spatial sign correlation compared to other nonparametric correlation estimators under equal marginal variances (left) and the influence function of the spatial sign correlation for $\rho = 0$ and $a = 1$ (right).

which is depicted in Figure 3.3 (left). Croux and Dehon (2010) compute the gross-error sensitivities of several nonparametric correlation measures at bivariate normal distributions. Figure 3.3 (left) corresponds to their Figure 2, complemented by the GES curve of the spatial sign correlation. The GES is small for any ρ, indicating a good robustness against small amounts of outliers. We refrain from stating the GES for arbitrary a and ρ explicitly since the formula is rather lengthy.

In Section 3.5, we will compare several correlation estimators with respect to their efficiency at the normal model. As a first glimpse in this direction, we recall the asymptotic variance of the Pearson correlation $\hat{\rho}_{\mathrm{Pea}}$ at elliptical distributions $ASV(\hat{\rho}_{\mathrm{Pea}}) = (1 + \kappa/3)(1 - \rho^2)^2$, where κ is the excess kurtosis of the components of F. The asymptotic relative efficiency of $\hat{\rho}_n$ with respect to $\hat{\rho}_{\mathrm{Pea}}$ is hence

$$ARE(\hat{\rho}_n, \hat{\rho}_{\mathrm{Pea}}) = \frac{ASV(\hat{\rho}_{\mathrm{Pea}})}{ASV(\hat{\rho}_n)} = \frac{1 + \kappa/3}{1 + \frac{1}{2}(a + a^{-1})(1 - \rho^2)^{-1/2}},$$

which is depicted in Figure 3.4 for normality, i.e. $\kappa = 0$. The maximum $1/2$ is attained for $a = 1$ and $\rho = 0$. If we fix $a = 1$, the asymptotic relative efficiency declines with increasing $|\rho|$, even tending to 0 for $|\rho| \to 1$. But it declines very slowly: for $|\rho| < 0.7$ it stays above 0.4. Under heavy-tailed distributions, however, the spatial sign correlation can be more efficient than the Pearson correlation. Specifically, $ARE(\hat{\rho}_n, \hat{\rho}_{\mathrm{Pea}}) \geq 1$ if $\kappa \geq (3/2)(a + a^{-1})/\sqrt{1 - \rho^2}$. For instance, with the kurtosis of the t_ν distribution being $6/(\nu - 4)$, the spatial sign correlation is more efficient at the bivariate spherical t_ν distribution for $\nu < 6$.

At any spherical distribution, i.e., for $a = 1$ and $\rho = 0$, the spatial sign correlation has an asymptotic variance of 2, which is the same as of the correlation estimator derived from Tyler's scatter matrix $\hat{V}_n$, cf. Section 3.5.3. However, it should be noted that in this case the asymptotic covariance matrix $ASV(\hat{V}_n)$ of the appropriately scaled Tyler matrix $\hat{V}_n$ (i.e., scaled such that $\mathrm{tr}(\hat{V}_n) = 1$) is $W_0/2$, which is four times the asymptotic covariance matrix of the SSCM. This is not a contradiction to the previous statement. The functions applied to the two matrix estimates to obtain a correlation estimate are different. Keep in mind that, when comparing expressions for the asymptotic variances of the SSCM and the Tyler matrix, the latter is usually scaled such that $\mathrm{tr}(\hat{V}_n) = p$ (as in Tyler, 1987) or $\det(\hat{V}_n) = 1$.

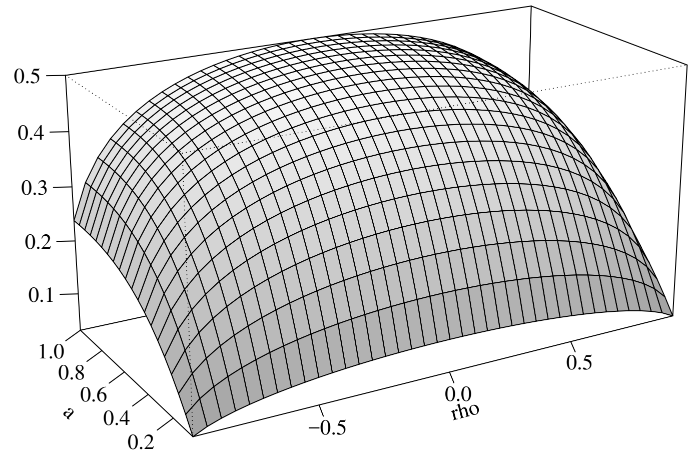

Figure 3.4: The asymptotic relative efficiency of $\hat{\rho}$ with respect to the empirical correlation under normality as a function of ρ and $a = \sqrt{v_{11}/v_{22}}$.

3.4 The two-stage spatial sign correlation

For fixed ρ, the asymptotic variance $ASV(\hat{\rho}_n)$ is minimized by equal marginal variances, that is, when $a = 1$. On the other hand, the asymptotic variance as well as the gross error sensitivity become arbitrarily large as a approaches 0 or ∞. It is therefore advisable to apply this estimator to standardized data, i.e. the components should be divided beforehand by a scale measure to yield equally dispersed margins. Margin-wise standardization generally should be administered with caution in multivariate data analysis, since it changes the shape, e.g., the direction of the eigenvectors, and will alter the results of, e.g., a principal component analysis. The inefficiency of the spatial sign covariance matrix at strongly "shaped", i.e. non-spherical, distributions has led to criticism regarding its use for robust principal component analysis, where a strong "shapedness" is the working assumption, cf. e.g. Remark 5.1 in Bali et al. (2011). We define shapedness as deviation from sphericity (and measure it for instance by the condition number of V). There are two sources that contribute to the shapedness: collinearity (ρ close to ± 1) and heteroscedasticity (a away from 1). The formula in Proposition 3 (2) nicely visualizes the individual influences of these two sources of shapedness on the asymptotic variance of $\hat{\rho}_n$. Since we are interested in correlation, a function of the shape that is invariant with respect to margin-wise scale changes, we can avoid the inefficiency due to the heteroscedasticity by margin-wise standardisation.

Let $\sigma(\cdot)$ denote a univariate *scale measure* or *dispersion measure*, i.e., for any univariate distribution G it satisfies

$$\sigma(G^{\star}_{\alpha,\beta}) = |\alpha|\, \sigma(G) \qquad \text{for all } \alpha, \beta \in \mathbb{R}, \tag{3.12}$$

where $G^{\star}_{\alpha,\beta}$ is the distribution of $Y^{\star} = \alpha Y + \beta$ for $Y \sim G$. This may be the standard deviation $\sigma_{SD} = \{E(Y - EY)^2\}^{1/2}$, but since the main purpose of studying spatial sign methods is their robustness, robust measures like the median absolute deviation $\sigma_{\mathrm{MAD}} = \mathrm{median}|Y - \mathrm{median}(Y)|$ or the Q_n scale measure $\sigma_{Q_n} = q_{1/4}(|Y - Y'|)$ may be more appropriate. Here, Y' is an independent copy of Y, median(Y) denotes the median of the distribution of Y, and $q_{1/4}(Y)$ its 1/4th quantile. Let further $\hat{\sigma}_n = \hat{\sigma}_n(\mathbb{Y}_n)$ denote the respective *scale estimator*, which is, in principle, the measure $\sigma(\cdot)$ applied to the empirical distribution associated with the univariate sample $\mathbb{Y}_n = (Y_1, \ldots, Y_n)^{\top}$. In many situations, the empirical version of the scale measure may be defined slightly differently due to various reasons, e.g., the empirical standard deviation is usually defined as $\hat{\sigma}_n(\mathbb{Y}_n) = \{(n-1)^{-1}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2\}^{1/2}$ instead of $\hat{\sigma}_n(\mathbb{Y}_n) = \{n^{-1}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2\}^{1/2}$, but the differences are negligible asymptotically.

Returning to the general p-dimensional set-up, for any specific choice of $\sigma(\cdot)$, let F_i denote the ith margin of F, further $\sigma_i = \sigma(F_i)$ and $\hat{\sigma}_{i,n} = \hat{\sigma}_n(\mathbb{X}_n^{(i)})$, where $\mathbb{X}_n^{(i)}$ is the ith column of $\mathbb{X}_n$, $1 \le i \le p$. Let

$$A = \begin{pmatrix} \sigma_1^{-1} & & 0 \\ & \ddots & \\ 0 & & \sigma_p^{-1} \end{pmatrix}, \qquad A_n = \begin{pmatrix} \hat{\sigma}_{1,n}^{-1} & & 0 \\ & \ddots & \\ 0 & & \hat{\sigma}_{p,n}^{-1} \end{pmatrix}.$$

Then we define the *two-stage spatial sign covariance matrix* as

$$\tilde{S}_n(\mathbb{X}_n, \boldsymbol{t}_n(\cdot), A_n) = \hat{S}_n(\mathbb{X}_n A_n, \boldsymbol{t}_n(\cdot)) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{s}\{A_n X_i - \boldsymbol{t}_n(\mathbb{X}_n A_n)\}\boldsymbol{s}\{A_n X_i - \boldsymbol{t}_n(\mathbb{X}_n A_n)\}^\top, \quad (3.13)$$

and the *two-stage spatial sign correlation* $\hat{\rho}_{\sigma,n}$ (of the sample $\mathbb{X}_n$ with location $\boldsymbol{t}_n(\cdot)$ and inverse scales A_n) as the spatial sign correlation $\hat{\rho}_n$, see (3.11), being applied to $\tilde{S}_n(\mathbb{X}_n, \boldsymbol{t}_n(\cdot), A_n)$ instead of $S_n(\mathbb{X}_n, \boldsymbol{t}_n)$.

Remark 1.

(I) *There is a subtle but important difference in the role that $\boldsymbol{t}_n$ plays in $\hat{S}_n(\mathbb{X}_n, \boldsymbol{t}_n)$ and in $\tilde{S}_n(\mathbb{X}_n, \boldsymbol{t}_n(\cdot), A_n)$. The location $\boldsymbol{t}_n$ may generally be any random vector, which may or may not bear a connection to the sample $\mathbb{X}_n$. But usually, we take it to be an estimator computed from the data, i.e., it is a function of $\mathbb{X}_n$. Whenever we want to invoke this latter meaning, we write $\boldsymbol{t}_n(\cdot)$ instead of $\boldsymbol{t}_n$, particularly so in the definition of $\tilde{S}_n(\mathbb{X}_n, \boldsymbol{t}_n(\cdot), A_n)$. Here it is essential that $\boldsymbol{t}_n(\cdot)$ is applied to the transformed data $\mathbb{X}_n A_n$.*

(II) *In the definition of the two-stage spatial sign covariance matrix $\tilde{S}_n(\mathbb{X}_n, \boldsymbol{t}_n(\cdot), A_n)$, the data are* first *standardized marginally, and* then *the location is estimated from the transformed data. For all marginally equivariant location estimators – and this is vast majority – the order of these two-steps is irrelevant. We call a multivariate location estimator $\boldsymbol{t}_n$* marginally equivariant *if it satisfies $\boldsymbol{t}_n(\mathbb{X}_n A + \boldsymbol{b}) = A\boldsymbol{t}_n(\mathbb{X}_n) + \boldsymbol{b}$ for any $p \times p$ diagonal matrix A and any $\boldsymbol{b} \in \mathbb{R}^p$. All location estimators being composed of univariate, affine equivariant location estimators are marginally equivariant. So are also all multivariate, affine equivariant location estimators, including elliptical maximum likelihood estimators, M-estimators (Maronna, 1976; Tyler, 1987), S-estimators (Davies, 1987), CM-estimators (Kent and Tyler, 1996), or MM-estimators (Tyler, 2002). However, there is one prominent example which lacks this property: the spatial median (e.g. Oja, 2010, Section 6.2). We want to include this estimator since, due to its conceptual similarity to the SSCM, it may be regarded as a default choice for $\boldsymbol{t}_n(\cdot)$. The spatial median has a variety of good properties such as uniqueness and computational and statistical efficiency, see e.g. Magyar and Tyler (2011) and the references therein. Likewise to the spatial sign covariance matrix, the spatial median is inefficient at strongly shaped distributions. Thus, when using the spatial median as location estimate, it is therefore reasonable to compute it from the marginally standardized data. This is the reason for choosing the order of steps as we do here: first standardization, then location estimation. However, in practical situations, the difference to the estimator obtained when reversing the order of these two steps tends to be rather small – also in case of the spatial median.*

(III) *Finally we would like to stress that we deliberately avoid any reference to the covariance matrix of F. Our whole discussion of scale and correlation is completely moment-free.*

Our main focus here is on estimating the generalized correlation coefficient ρ within the semiparametric model of elliptical distributions, but the concept of spatial sign correlation can also be employed for defining a general, moment-free measure of association. Requiring no moment assumptions is one major strength of spatial sign methods.

3.4.1 Asymptotic results

The first result concerns the asymptotic difference between $\hat{S}_n(\mathbb{X}_n A_n, \boldsymbol{t}_n(\cdot))$, the sample two-stage SSCM with estimated location and scales, and $\hat{S}_n(\mathbb{X}_n A, A\boldsymbol{t})$, the sample two-stage SSCM with known location and scales. We use the notation $X^{(j)}$ to denote the jth component of the p-dimensional random vector $\boldsymbol{X}$, $j = 1, \ldots, p$, likewise for other vectors.

Theorem 1. *(Dürre and Vogel, 2016a) Let $\boldsymbol{t} \in \mathbb{R}^p$ and $\boldsymbol{X}$ be a p-variate random vector with continuous distribution F satisfying*

(C1) $\mathrm{E}|\boldsymbol{X} - \boldsymbol{t}|^{-3/2} < \infty$,

(C2) $\mathrm{E}\left\{\frac{\boldsymbol{X}-\boldsymbol{t}}{|\boldsymbol{X}-\boldsymbol{t}|^2}\right\} = 0$ *and* $\mathrm{E}\left\{\frac{(\boldsymbol{X}-\boldsymbol{t})^{(i)}(\boldsymbol{X}-\boldsymbol{t})^{(j)}(\boldsymbol{X}-\boldsymbol{t})^{(k)}}{|\boldsymbol{X}-\boldsymbol{t}|^4}\right\} = 0$ *for* $i, j, k = 1, \ldots, p$.

Let further A be a $p \times p$ diagonal matrix with positive diagonal entries $a_1, \ldots, a_p$, and A_n a series of random $p \times p$ diagonal matrices satisfying

(C3) $\sqrt{n}(A_n - A) \xrightarrow{d} Z = \mathrm{diag}(Z_1, \ldots, Z_p)$

for some random diagonal matrix Z. Finally, let $\mathbb{X}_n = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^\top$ be an iid sample drawn from F and $\boldsymbol{t}_n(\cdot)$ a series of p-variate estimators satisfying

(C4) $\sqrt{n}\{\boldsymbol{t}_n(\mathbb{X}_n) - \boldsymbol{t}\} = O_P(1)$,

(C5) $\sqrt{n}\{\boldsymbol{t}_n(\mathbb{X}_n A_n) - A_n \boldsymbol{t}_n(\mathbb{X}_n)\} = O_P(1)$.

Then $\sqrt{n}\{S_n(\mathbb{X}_n A_n, \boldsymbol{t}_n(\cdot)) - S_n(\mathbb{X}_n A, A\boldsymbol{t})\} \xrightarrow{d} \Xi_p$ *as $n \to \infty$ with*

$$\Xi_p = A^{-1} Z S(F_0, 0) + S(F_0, 0) A^{-1} Z - 2 \sum_{j=1}^{p} (Z_j / a_j) \Gamma_j, \tag{3.14}$$

where F_0 is the distribution of $\boldsymbol{X}_0 = A(\boldsymbol{X} - \boldsymbol{t})$ and

$$\Gamma_j = \mathrm{E}\left\{ \left(X_0^{(j)}\right)^2 \frac{\boldsymbol{X}_0 \boldsymbol{X}_0^\top}{\left(\boldsymbol{X}_0^\top \boldsymbol{X}_0\right)^2} \right\}.$$

Theorem 1 apparently has a long list of technical conditions. They are due to the fact that it is formulated under very broad conditions. We do not assume any specific model for the distribution F. Also, the location estimator $\boldsymbol{t}_n(\cdot)$, the scale estimator A_n and even the location $\boldsymbol{t}$ are unspecified. The above conditions are indeed a set of easy-to-verify regularity conditions, which are met in practically all relevant situations, and many of which may be further relaxed for the price of more involved technical derivations. We will review them one by one below.

Condition (C1) requires the probability mass of F to be not too strongly concentrated around $\boldsymbol{t}$. For instance, if F possesses a Lebesgue density f, it is sufficient (but not necessary) that f is bounded at $\boldsymbol{t}$. This condition also appears in Theorems 2 and 3 of Dürre et al. (2014) and is, loosely speaking, due to the discontinuity of the spatial sign function at the origin.

Condition (C2) is indeed a somewhat restrictive condition as it basically imposes componentwise symmetry of F around $\boldsymbol{t}$. It is, however, a mere convenience assumption, it can be dropped in favor of an additional term in (3.14) and a slightly stronger formulation of the other conditions (basically joint convergence of $\hat{S}_n$, $\boldsymbol{t}_n$ and A_n). The proof of the more general version runs analogously, with the main difference that Dürre et al. (2014, Theorem 3) instead of Dürre et al. (2014, Theorem 2) is used. However, our central result, Theorem 2 below, concerns elliptical distributions, for which (C2) is fulfilled. We therefore consider it appropriate to include this symmetry condition here for the sake of simpler conditions and a clearer exposition.

Condition (C3) is satisfied, e.g., if A_n^{-2} is taken to be the diagonal of some $p \times p$ scatter matrix estimator for which asymptotic normality has been shown. But also if A_n^{-1} is composed of univariate scale estimators (the default case here due to computational reasonability), it is usually true. Specifically, if the univariate scale estimator $\hat{\sigma}_{j,n}$ allows a linearization, i.e.,

$$\hat{\sigma}_{j,n} = \frac{1}{n}\sum_{i=1}^{n} f_j(X_i^{(j)}) + o_p(n^{-1/2}), \qquad j = 1, \ldots, p, \tag{3.15}$$

with $\mathrm{E}\{f_j(X^{(j)})^2\} < \infty$, then $\sqrt{n}\{(\hat{\sigma}_{1,n}, \ldots, \hat{\sigma}_{p,n}) - (\sigma_1, \ldots, \sigma_p)\}^\top = \sqrt{n}\,\mathrm{diag}(A_n^{-1} - A^{-1})$ converges to a multivariate normal distribution, and then so does $\sqrt{n}(A_n - A)$. Note that, since A and A_n are diagonal matrices, $\sqrt{n}(A_n^{-1} - A^{-1}) \xrightarrow{d} \tilde{Z}$ implies $\sqrt{n}(A_n - A) = AA_n\sqrt{n}(A^{-1} - A_n^{-1}) \xrightarrow{d} -A^2\tilde{Z}$, and hence $Z = -A^2\tilde{Z}$ in distribution.

All estimators of practical relevance allow a linearization (3.15). For instance, for quantile-based estimators, such as the MAD, this linearization is provided by the Bahadur representation (Bahadur, 1966; Kiefer, 1967; Ghosh, 1971; Sen, 1968). In the case of U-statistics, such as Gini's mean difference, it is given by the Hoeffding decomposition (Hoeffding, 1948), and in the case of U-quantiles, such as the Q_n scale estimator (Rousseeuw and Croux, 1993), by a combination of the two (Serfling, 1984; Wendler, 2011).

Condition (C4): This is a minimal standard assumption.

Condition (C5) is trivially fulfilled for any marginally equivariant location estimator, see Remark 1 (II). This condition is necessary since we want to include the spatial median as potential location estimator, and, for efficiency reasons, propose to standardize the data prior to computing its spatial median – instead of scaling the spatial median along with the data. Under (C3), the spatial median satisfies (C5) at elliptical distributions (Nevalainen et al., 2007).

Finally, the continuity of F also is a mere convenience assumption, which prohibits that several data points coincide with each other, and thus ensures that $\boldsymbol{t}_n$ coincides with at most one observation. Alternative assumptions are discussed also in Dürre et al. (2014).

In case of F being an elliptical distribution and $\boldsymbol{t}$ its symmetry center, explicit expressions for $S(F, \boldsymbol{t})$ appear to be known only for $p = 2$. In this case, Ξ_p in (3.14) considerably simplifies.

Corollary 1. *Let $p = 2$ and $\boldsymbol{X} \sim F \in \mathscr{E}_2(\boldsymbol{t}, V)$. Let $A = \mathrm{diag}(a_1, a_2)$ be a 2×2 diagonal matrix with positive diagonal entries such that $V_0 = AVA$ has equal diagonal entries. Denote the diagonal entries of the diagonal matrix Z from Theorem 1 by Z_1 and Z_2. Then Ξ_2 from Theorem 1 is*

$$\Xi_2 = \begin{pmatrix} Z_1/a_1 - Z_2/a_2 & 0 \\ 0 & Z_2/a_2 - Z_1/a_1 \end{pmatrix} \zeta,$$

where $\zeta = (1 - \sqrt{1 - \rho^2})/(2\rho^2)$ if $\rho \neq 0$ and $\zeta = 1/4$ if $\rho = 0$, and $\rho = v_{12}(v_{11}v_{22})^{-1/2}$.

An important implication of Corollary 1 is that, at elliptical population distributions, the asymptotic distribution of the off-diagonal element of the two-dimensional two-stage SSCM is the same as that of the off-diagonal element of the ordinary SSCM at the corresponding distribution with equal marginal scales. Building on this observation, we can derive the asymptotic distribution of the two-stage spatial sign correlation by means of a generalized version of the delta method.

Theorem 2. *(Dürre and Vogel, 2016a) Let $p = 2$ and $\boldsymbol{X} \sim F \in \mathscr{E}_2(\boldsymbol{t}, V)$ satisfy Condition (C1) of Theorem 1. Let $\mathbb{X}_n$, A, A_n and $\boldsymbol{t}_n(\cdot)$ be as in Theorem 1, satisfying Conditions (C3), (C4) and (C5), with the further property that $V_0 = AVA$ has equal diagonal entries. Then*

$$\sqrt{n}(\hat{\rho}_{\sigma,n} - \rho) \ \xrightarrow{d}\ \mathcal{N}\left\{0, (1-\rho^2)^2 + (1-\rho^2)^{3/2}\right\}. \tag{3.16}$$

We have the following remarks about Theorem 2.

Remark 2.

(I) *Comparing Theorem 2 to Proposition 3 (2), we find that, at any elliptical distribution, the spatial sign correlation with the margins being standardized beforehand by the* true *scales and the spatial sign correlation with the margins being standardized by* estimated *scales have the same asymptotic efficiency. In fact, we show in the Appendix that they are asymptotically equivalent. In other words, the loss for not knowing the scale is zero asymptotically, and this is true regardless of the scale estimator used. Any scale function $\sigma(\cdot)$ satisfying (3.12) yields that $\boldsymbol{X}_0 = A\boldsymbol{X}$ has equal marginal scales if $\boldsymbol{X}$ is elliptical. Also, the finite-sample variances of the spatial sign correlation with known and estimated scales hardly differ, as the simulations in Subsection 3.4.2 indicate.*

(II) *At elliptical distributions with finite fourth moments, the asymptotic variance of the product moment correlation is $(1+\kappa/3)(1-\rho^2)^2$, where κ is the marginal excess kurtosis. Thus under normality, where $\kappa = 0$, the additional term $(1-\rho^2)^{3/2}$ may be viewed as the price to pay efficiency-wise for the gain in robustness when using the two-stage spatial sign correlation instead of the moment correlation.*

(III) *Note that in case of a two-dimensional elliptical distribution, Condition (C2) from Theorem 1 is always fulfilled and Condition (C1) if $g(z) = O(z^{-1/4+\delta})$ as $z \to 0$ for some $\delta > 0$.*

The asymptotic distribution of $\hat{\rho}_{\sigma,n}$ only depends on ρ, but not on the elliptical generator g or any other characteristic of the population distribution. Therefore the two-stage spatial sign correlation is very well suited for nonparametric and robust correlation testing. Likewise to Fisher's z-transformation for the moment correlation under normality (Fisher, 1921), one can find a variance-stabilizing transformation for the spatial sign correlation under ellipticity.

Corollary 2. *(Dürre and Vogel, 2016a) Under the conditions of Theorem 2, we have $\sqrt{n}\{h(\hat{\rho}_{\sigma,n}) - h(\rho)\} \xrightarrow{d} \mathcal{N}(0,1)$ with*

$$h(x) = s(x)\left[\frac{1}{\sqrt{2}}\arcsin\left\{\frac{3(1-\sqrt{1-x^2})-2}{\sqrt{1-x^2}+1}\right\} + \frac{\pi}{2^{3/2}}\right], \tag{3.17}$$

where $s(\cdot)$ denotes the (in this case univariate) sign function.

As can be seen in Figure 3.5, the transformation h is similar to Fisher's z-transform $z(x) = \log\{(1+x)/(1-x)\}/2$. There are two main differences: first, h is flatter, with a smaller derivative throughout, reflecting the larger asymptotic variance of the spatial sign correlation

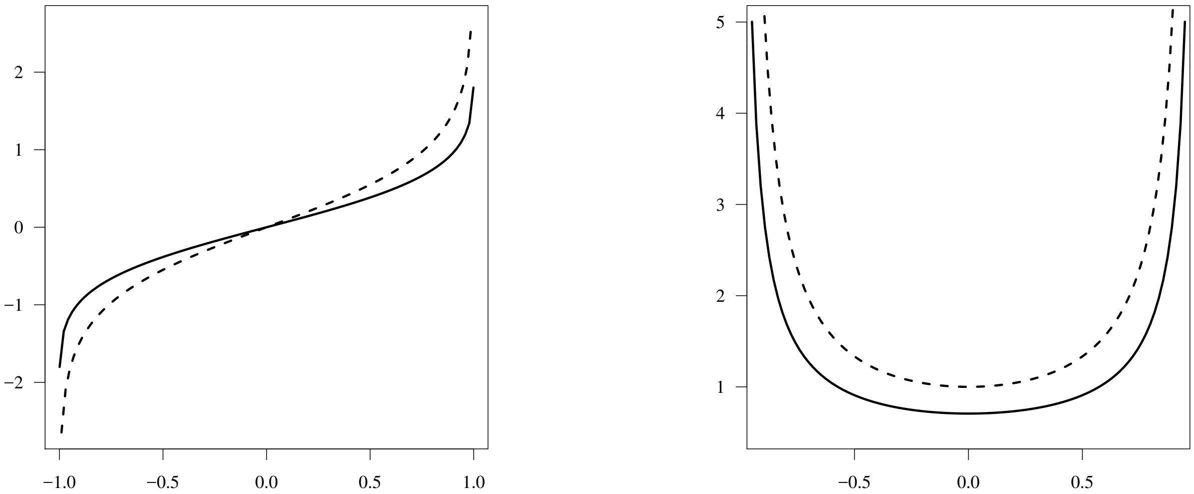

Figure 3.5: Variance-stabilizing transformations (left) and their derivates (right) for the spatial sign correlation (solid) and the Pearson moment correlation, i.e., Fisher's z-transform (dashed).

under normality, and second, h is bounded, attaining only values between $-\pi/\sqrt{2}$ and $\pi/\sqrt{2}$. To construct confidence intervals, its inverse function $h^{-1} : [-\pi/\sqrt{2}, \pi/\sqrt{2}] \to [-1, 1]$ is also of interest. It is given by

$$h^{-1}(y) = s(y) \frac{2^{3/2}\sqrt{1-\cos(\sqrt{2}y)}}{3-\cos(\sqrt{2}y)}.$$

Based on Corollary 2, one can derive asymptotic level-α-tests for the generalized correlation coefficient ρ of a bivariate elliptical distribution, which are robust and very accurate also in small samples, as the results of Section 3.4.2 below indicate. For instance, a two-sided one-sample test for ρ based on $\hat{\rho}_{\sigma,n}$ would reject the null hypothesis $\rho = \rho_0$ at the significance level α if the test statistic

$$T_{1,n} = n\{h(\hat{\rho}_{\sigma,n}) - h(\rho_0)\}^2$$

exceeds $\chi^2_{1;1-\alpha}$, i.e., the $1-\alpha$ quantile of the χ^2 distribution with one degree of freedom. Likewise, for two samples of sizes n_1 and n_2 and generalized correlation coefficients $\rho^{(1)}$ and $\rho^{(2)}$, respectively, the null hypothesis $\rho^{(1)} = \rho^{(2)}$ is rejected if

$$T_{2,n} = \frac{n_1 n_2}{n_1 + n_2}\{h(\hat{\rho}^{(1)}_{\sigma,n}) - h(\hat{\rho}^{(2)}_{\sigma,n})\}^2$$

is larger than $\chi^2_{1;1-\alpha}$, where $\hat{\rho}^{(i)}_{\sigma,n}$, $i = 1, 2$, denote the two-stage spatial sign correlations computed from the two samples. The asymptotic χ^2_1 distribution of the two-sample test statistic is derived under the assumption that $n_1, n_2 \to \infty$ and $n_1/(n_1 + n_2) \to \lambda \in (0, 1)$. Similarly, one can construct one-sided and k-sample tests.

3.4.2 Simulations

First we examine the influence of the chosen scale estimator in finite samples, respectively the difference between knowing and estimating the scale at all. Results for a bivariate normal distribution with $a = 10$ and $\rho = 0$ are presented in Table 3.1. We compare the n-stabilized variances (based on 100,000 replications for each n) of the two-stage spatial sign correlation

	$\hat{\rho}_{\sigma,n}$ (two-stage)				$\hat{\rho}_n$
n	known	SD	Q_n	MAD	
10	1.703	1.664	1.684	1.706	3.953
20	1.815	1.795	1.799	1.804	4.688
50	1.921	1.916	1.914	1.913	5.429
100	1.936	1.930	1.928	1.926	5.614
500	1.973	1.972	1.971	1.972	5.945

Table 3.1: Simulated variances (based on 100,000 replications), multiplied by the sample size n, of the ordinary and the two-stage spatial sign correlation at a bivariate normal distribution with $a = 10, \rho = 0$ and different sample sizes n. Standardization of the two-step estimator is by the true marginal standard deviation (known), by the sample standard deviation (SD), the Q_n and the MAD.

(pre-standardized by the standard deviation, the Q_n and the MAD), the one-step spatial sign correlation and the one-step spatial sign correlation being applied to the data standardized by the true scale parameter. We observe that there is a large improvement through the pre-standardization, and that it makes practically no difference, also for small n, if the true scale is known or has to be estimated, and if so, which scale estimator is used. Results for other distributions, a and ρ are comparable. Standardizing by the standard deviation is, as expected, slightly worse at very heavy tailed distributions.

Furthermore we want to numerically investigate the usefulness of the asymptotics and especially the h-transform (3.17) in finite samples. We compute 95% confidence intervals based on the spatial sign correlation without and with the transformation h, denoted in the tables below by sscor and sscor-h, respectively. The confidence intervals without transformation are given by

$$\hat{\rho}_{\sigma,n} \pm \sqrt{\frac{(1-\hat{\rho}_{\sigma,n}^2)^2 + (1-\hat{\rho}_{\sigma,n}^2)^{3/2}}{n}} z_{1-\alpha/2}$$

and with transformation by

$$h^{-1}\left\{h(\hat{\rho}_{\sigma,n}) \pm \frac{z_{1-\alpha/2}}{\sqrt{n}}\right\},$$

where $z_{1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of the standard normal distribution, which is $z_{0.975} = 1.96$ in our case. Pearson's moment correlation $\hat{r}_n$ serves as benchmark. Under ellipticity, the asymptotic variance of $\hat{r}_n$ additionally depends on the kurtosis κ. We estimate the latter by the following multivariate kurtosis estimator

$$\hat{\kappa}_n = \frac{3}{p(p+2)} \frac{1}{n} \sum_{i=1}^{n} \{(X_i - \bar{X}_n)^\top \hat{\Sigma}_n^{-1} (X_i - \bar{X}_n)\}^2 \; - \; 3,$$

where $\bar{X}_n$ denotes the sample mean and $\hat{\Sigma}_n$ the sample covariance matrix (e.g. Anderson, 2003, p. 103). Alternatively, one may estimate the kurtosis by averaging the componentwise marginal sample kurtoses, as it is done, e.g., in Vogel and Fried (2011). The confidence intervals for ρ based on the sample moment correlation without and with z-transformation (denoted by cor and cor-z, respectively) are then given by

$$\hat{r}_n \pm \sqrt{\frac{1+\hat{\kappa}_n}{n}}(1-\hat{r}_n^2) z_{1-\alpha/2} \qquad \text{and} \qquad z^{-1}\left\{z(\hat{r}_n) \pm \sqrt{\frac{1+\hat{\kappa}_n}{n-3}} z_{1-\alpha/2}\right\},$$

respectively, where $z(\cdot)$ denotes the z-transform. The simulations are done with the statistical software R (R Core Team, 2016). We sample from bivariate elliptical distributions using the package mvtnorm (Genz et al., 2014). The central location is computed by the spatial median from the package pcaPP (Filzmoser et al., 2011), and the scales are estimated by the Q_n implemented in the package robustbase (Rousseeuw et al., 2014).

In Table 3.2, covering frequencies of the generalized correlation coefficient ρ by the various confidence intervals are given based on 10,000 repetitions for each parameter setting. We consider the normal distribution and the t-distribution with 5 and 3 degrees of freedom, true correlations of $\rho = 0$ and $\rho = 0.5$ and six different sample sizes ranging from $n = 10$ to $n = 10,000$.

We see that the sscor-h confidence intervals, i.e., the spatial-sign-based with transformation h, are almost exact in all cases considered, already for $n = 10$. The sscor-based intervals without transformation as well as the z-transformed $\hat{r}_n$-based intervals reach a comparable accuracy only for $n = 50$, and the $\hat{r}_n$-based confidence intervals without z-transformation no sooner than $n = 100$ even under normality. Table 3.3 reports the corresponding average lengths of the confidence intervals multiplied by $\sqrt{n}$. Comparing these average lengths for the Pearson correlation and spatial sign correlation, we rediscover roughly the square root of the ratio of the asymptotic variances, e.g., for the normal distribution at $\rho = 0$, we have $5.54/3.92 = 1.413 \approx \sqrt{2}$. At normality, the confidence intervals based on the Pearson correlation (the maximum likelihood estimator for ρ in this case) are shorter, whereas the sscor confidence intervals are shorter at the t_5 distribution – at least in larger samples, where all confidence intervals have the same 95% covering probability. Thus, in a heavy-tailed setting like the t_5 distribution, the spatial-sign-based confidence intervals are superior in terms of covering accuracy as well as length. Further, we observe that the strict asymptotic distribution-freeness of the spatial sign correlation practically also extends to the finite-sample case. In both tables, the results for the spatial sign correlation are essentially the same for the three different elliptical distributions. In contrast, the Pearson correlation shows a considerably worse finite-sample behavior at the t_5 than at the normal distribution.

It should also be noted that, when constructing confidence intervals using Fisher's z-transform, we have employed that $\text{var}\{z(\hat{r}_n)\} \approx 1/(n-3)$ yields a better approximation than $\text{var}\{z(\hat{r}_n)\} \approx 1/n$. It appears that the appropriate standard error for the h-transform of the spatial sign correlation is $1/\sqrt{n}$. If the z-transform is applied naively, i.e., with $1/\sqrt{n}$ standard errors, the accuracy comparison of the confidence intervals is yet much more favorable for the spatial sign correlation.

When sampling from a t_3 distribution, where fourth moments are not finite, the usual construction of the moment-correlation-based confidence intervals lacks a mathematical justification. However, the bottom parts of Tables 3.2 and 3.3 indicate that they nevertheless provide a somewhat useful approximation. While for small n the moment-correlation-based confidence intervals are short but have a too low coverage probability, they reach 95% in large samples, but are in comparison to the sscor-based confidence intervals very large. The slower convergence of the sample moment correlation to ρ, and the exploding behavior of the sample kurtosis are opposing effects, which appear to basically cancel each other.

Altogether the spatial correlation with variance stabilizing-transformation h yields very reliable confidence bands, which are accurate also in very small samples.

3.5 Comparison of robust correlation estimators

There are many proposals for robust correlation estimators in the literature. In this section we compare the spatial sign correlation $\hat{\rho}_n$ to a number of prominent alternatives, without claiming or attempting any completeness or ranking. In Subsections 3.5.1-3.5.3, we gather the basic analytic results, particularly the asymptotic efficiencies at the normal model, and in Subsection

ρ	0				0.5			
n	sscor	sscor-h	cor	cor-z	sscor	sscor-h	cor	cor-z
normal distribution								
10	86	94	77	89	87	93	78	89
20	90	94	86	92	91	95	87	92
50	93	95	92	94	93	95	92	93
100	93	95	93	94	94	95	93	94
500	95	95	95	95	95	95	95	95
10000	95	95	95	95	95	95	95	95
t_5 distribution								
10	85	94	70	84	87	93	71	84
20	90	95	81	88	90	95	80	88
50	93	95	88	91	93	95	88	91
100	94	95	91	93	94	95	91	93
500	95	95	94	94	95	95	94	94
10000	95	95	95	95	95	95	95	95
t_3 distribution								
10	85	94	64	80	87	93	66	80
20	90	94	74	83	90	94	76	84
50	93	95	82	87	93	95	82	86
100	94	95	86	89	94	95	86	89
500	95	95	90	92	94	95	90	92
10000	95	95	94	95	95	95	94	94

Table 3.2: Empirical covering probabilities (%) of asymptotic 95% confidence intervals based on the spatial sign correlation (sscor) and the moment correlation (cor) with and without variance-stabilizing transformation for bivariate normal and t-distributions with 3 and 5 degrees of freedom, $\rho = 0$ and $\rho = 0.5$, and varying sample sizes n; 10,000 repetitions.

3.5.4 we compare the finite-sample and robustness properties numerically.

In general, the estimators mentioned are known to be Fisher-consistent for the correlation only under normality, which often, as in the case of the spatial sign correlation, can be relaxed to ellipticity. To put it differently, each of the various correlation estimators $\hat{\theta}_n$ estimates some parameter θ of the bivariate population distribution, which may serve as a measure of monotone dependence, but does in general not coincide with the moment correlation ρ.[3] The exact functional connection between θ and ρ is usually hard to assess for arbitrary distributions, but is known for the normal model. If no such function is mentioned in the examples below, it is the identity.

Let the data be denoted by $\boldsymbol{X}_i = (X_i, Y_i)^T$, $i = 1, \ldots, n$, independent and normally distributed. Relative efficiencies reported below are with respect to the sample correlation, which is denoted by $\hat{\rho}_{\mathrm{Pea}}$. The estimators we will consider can roughly be divided into three groups: We call the first group *nonparametric estimators* since they depend on signs and ranks. Besides the spatial sign correlation, these are the Gaussian rank correlation, Spearman's ρ, Kendall's τ, and the quadrant correlation. The second group are the *Gnanadesikan-Kettenring-type estimators*, where we consider the τ-scale and the Q_n as scale estimators. We label the third group *affine equivariant estimators*, i.e., estimators that are derived from an affine equivariant two-dimensional scatter estimator. Here we consider Tyler's M-estimator, the raw and the re-

[3] In this sense, "correlation" is understood as monotone dependence.

ρ	0				0.5			
n	sscor	sscor-h	cor	cor-z	sscor	sscor-h	cor	cor-z
normal distribution								
10	4.75	4.11	2.83	3.11	4.08	3.69	2.23	2.56
20	5.11	4.68	3.35	3.45	4.20	3.99	2.57	2.73
50	5.36	5.15	3.68	3.71	4.26	4.18	2.79	2.85
100	5.45	5.33	3.80	3.81	4.30	4.26	2.87	2.90
500	5.52	5.50	3.89	3.90	4.31	4.30	2.92	2.93
10000	5.54	5.54	3.92	3.92	4.32	4.32	2.94	2.94
t_5 distribution								
10	4.75	4.12	2.84	3.12	4.08	3.69	2.28	2.61
20	5.11	4.68	3.63	3.71	4.22	4.01	2.83	2.99
50	5.36	5.15	4.44	4.42	4.28	4.20	3.40	3.46
100	5.45	5.34	4.92	4.89	4.30	4.26	3.74	3.77
500	5.52	5.50	5.71	5.69	4.31	4.31	4.29	4.29
10000	5.54	5.54	6.38	6.38	4.32	4.31	4.79	4.79
t_3 distribution								
10	4.73	4.10	2.81	3.10	4.09	3.70	2.27	2.60
20	5.11	4.68	3.77	3.84	4.22	4.01	2.99	3.15
50	5.36	5.15	5.08	5.01	4.28	4.20	3.94	3.99
100	5.45	5.34	6.15	6.04	4.30	4.26	4.72	4.73
500	5.52	5.50	9.12	8.98	4.31	4.31	6.90	6.87
10000	5.54	5.54	17.57	17.46	4.32	4.32	13.02	12.99

Table 3.3: Average lengths of 95% confidence intervals based on the spatial sign correlation (sscor) and the moment correlation (cor) with and without variance-stabilizing transformation for bivariate normal and t-distributions with 3 and 5 degrees of freedom, $\rho = 0$ and $\rho = 0.5$, and varying sample sizes n; 10,000 repetitions.

weighted MCD, and the S-estimator with Tukey's biweight-function. Detailed descriptions of the estimators are given below.

3.5.1 Nonparametric estimators

The Gaussian rank correlation is defined as the sample correlation of the normal scores of the data, i.e.

$$\hat{\rho}_{\text{GRK}} = \frac{1}{c_n} \sum_{i=1}^{n} \Phi^{-1}\left(\frac{R(X_i)}{n+1}\right) \Phi^{-1}\left(\frac{R(Y_i)}{n+1}\right),$$

where $c_n = \sum_{i=1}^{n} \Phi^{-1}\left(\frac{i}{n+1}\right)^2$, $R(X_i)$ is the rank of X_i among $X_1, \ldots, X_n$, and Φ^{-1} is the quantile function of the standard normal distribution. The influence function of the Gaussian rank correlation is unbounded, but in finite samples it is much more robust than the Pearson correlation (Boudt et al., 2012). Since the Gaussian rank correlation corresponds to the Pearson correlation of the transformed data, the pairwise estimation of a multidimensional correlation matrix leads always to a non-negative definite estimate.

Another rank based estimator is Spearman's ρ, which is the sample correlation of the ranks $R(X_1), \ldots, R(X_n)$ and $R(Y_1), \ldots, R(Y_n)$. To obtain a consistent estimator for ρ, one has to apply the transformation $\hat{\rho}_{\text{Sp.c}} = 2\sin(\pi\hat{\rho}_{\text{Sp}}/6)$, which goes back to Pearson (1907). A further popular nonparametric estimator is Kendall's τ, which is defined as

$$\hat{\rho}_{\text{Ken}} = \frac{2}{n(n-1)} \sum_{i>j} \boldsymbol{s}\left((X_i - X_j)(Y_i - Y_j)\right),$$

where $\boldsymbol{s}(\cdot)$ is the sign function defined at the beginning of Section 3.2, here applied to a univariate argument. It also requires a consistency transformation, which is valid under ellipticity (e.g. Möttönen et al., 1999): $\hat{\rho}_{\text{Ken.c}} = \sin(\pi\hat{\rho}_{\text{Ken}}/2)$. A highly robust, non-parametric procedure based on signs is the quadrant correlation, which first appears in Mosteller (1946) and Blomqvist (1950). It can be expressed as $\hat{\rho}_{\text{Q}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{s}\left((X_i - \text{med}(X))(Y_i - \text{med}(Y))\right)$, where $\text{med}(X)$ denotes the median of $X_1, \ldots, X_n$. The same transformation $\hat{\rho}_{\text{Q.c}} = \sin(\pi\hat{\rho}_{\text{Q}}/2)$ renders this estimator consistent for ρ under elliptical distributions. All three nonparametric estimators $\hat{\rho}_{\text{Sp.c}}$, $\hat{\rho}_{\text{Ken.c}}$, $\hat{\rho}_{\text{Q.c}}$ have a bounded influence function and are therefore called B-robust. Their influence functions, asymptotic variances and gross-error sensitivities can be found in Croux and Dehon (2010).

3.5.2 GK estimators

Gnanadesikan and Kettenring (1972) introduced an estimation principle based on robust variance estimation,

$$\hat{\rho} = \frac{\hat{\sigma}^2(X/\sigma_1 + Y/\sigma_2) - \hat{\sigma}^2(X/\sigma_1 - Y/\sigma_2)}{\hat{\sigma}^2(X/\sigma_1 + Y/\sigma_2) + \hat{\sigma}^2(X/\sigma_1 - Y/\sigma_2)},$$

where $\hat{\sigma}$ can be any robust scale measure and $\sigma_1 = \hat{\sigma}(X)$, $\sigma_2 = \hat{\sigma}(Y)$. Such an estimator can be seen to be Fisher-consistent for ρ, regardless of the choice of the scale measure $\hat{\sigma}$, if $X+Y$, $X-Y$ as well as X and Y have the same distribution up to location and scale, which is fulfilled for elliptical distributions. According to Ma and Genton (2001), the correlation estimator has the same asymptotic relative efficiency (with respect to the Pearson correlation) at the normal distribution as the underlying scale estimator posses with respect to the standard deviation. There is also a relationship between the influence functions, which guarantees that the B-robustness translates from the variance to the correlation estimator, see Genton and Ma (1999). In the recent literature, there are two proposals for the variance estimation. Maronna and Zamar (2002) favor the so-called τ-scale (not to be confused with Kendall's τ):

$$\hat{\sigma}_\tau = \frac{\sigma_0^2}{n} \sum_{i=1}^{n} d_{c_2}\left(\frac{X_i - \hat{\mu}(X)}{\sigma_0}\right), \qquad \text{where} \qquad \hat{\mu}(X) = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i},$$

$w_i = W_{c_1}\{(X_i - \text{med}(X))/\sigma_0\}$, $\sigma_0 = \text{med}\{\,|X_i - \text{med}\,X| \; : \; i = 1, \ldots, n\}$, $d_c(x) = \min(x^2, c^2)$ and $W_c(x) = \{1 - (x/c)^2\}^2 \mathbb{1}_{\{|x| \le c\}}$. They use $c_1 = 4.5$ and $c_2 = 3$ to get an efficiency of approximately 0.8 under normality. Ma and Genton (2001) use the Q_n, which is defined as

$$Q_n(X) = d \cdot \{|X_i - X_j| : i < j\}_{(k)},$$

where $k = \binom{[n/2]+1}{2}$ and the notation $\{\cdot\}_{(k)}$ refers to the kth order statistic. The consistency factor d equals $1/(\sqrt{2}\Phi^{-1}(5/8))$ for the normal distribution. This estimator has an asymptotic efficiency of 0.82, see Rousseeuw and Croux (1993). The influence function of the resulting covariance estimator is bounded and can also be found in Ma and Genton (2001).

3.5.3 Affine equivariant estimators

One can estimate the correlation by means of any affine equivariant, bivariate scatter estimator $\hat{V}_n$ using the relation $\hat{\rho} = \hat{v}_{1,2}/\sqrt{\hat{v}_{1,1}\hat{v}_{2,2}}$. Taskinen et al. (2006) derive the influence function of the correlation estimator from the influence function of $\hat{V}_n$ under elliptical distributions. Furthermore, the asymptotic variance of $\hat{\rho}$ is of the form $(1-\rho^2)^2 \cdot ASV(\hat{v}_{1,2})$, where $ASV(\hat{v}_{1,2})$ is the asymptotic variance of $\hat{v}_{1,2}$ under the corresponding spherical distribution. We consider three examples of robust affine equivariant scatter estimators.

Tyler (1987) proposed an M-estimator for the shape matrix V, being a suitably scaled solution of

$$\frac{2}{n}\sum_{i=1}^{n} \frac{(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_n)(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_n)^T}{(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_n)^T \hat{V}_n^{-1}(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_n)} = \hat{V}_n,$$

where $\hat{\boldsymbol{\mu}}_n$ is a suitable multivariate location estimate. In the simulations in Section 3.5.4, we take the spatial median. The Tyler estimator can be regarded as an affine equivariant version of the SSCM and is also distribution-free within the elliptical model.

A highly robust, affine equivariant scatter estimator is the minimum covariance determinant (MCD) estimator proposed by Rousseeuw (1985). For a given trimming constant α, it is defined as the sample covariance matrix of the subset of observations which yields the smallest determinant of the matrix estimate among all subsets of size $\lfloor (1-\alpha)\cdot n \rfloor$. Choosing $\alpha = 0.5$ results in an asymptotic breakdown point of 0.5. Since the asymptotic efficiency, especially in small dimensions, is rather low, the raw MCD is usually followed by a reweighting step. We call this two-step estimate the weighted MCD. For both, the raw and the weighted MCD, influence functions, consistency factors and asymptotic efficiencies can be found in Croux and Haesbroeck (1999).

Davies (1987) proposed a multivariate generalization of S-estimators, being defined as

$$(\hat{\boldsymbol{\mu}}_n, \hat{V}_n) = \underset{\boldsymbol{\mu}, V}{\arg\min} \det(V) \qquad \text{subject to} \quad \operatorname*{ave}_{i=1}^{n} w(\hat{d}_i) = b,$$

where $d_i = \{(\boldsymbol{X}_i - \boldsymbol{\mu})^T V^{-1}(\boldsymbol{X}_i - \boldsymbol{\mu})\}^{1/2}$ and w is a suitable, smooth and bounded, weight function, usually the Tukey–biweight:

$$w_c(y) = \min\left(\frac{y^2}{2} - \frac{y^4}{2c^2} + \frac{y^6}{6c^4}, \frac{c^2}{6}\right).$$

Letting $b = E\{w_c(\|V^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})\|)\}$ yields Fisher-consistency at the elliptical population distribution F, and if c is chosen such that $rc^2/6 = E\{w_c(\|V^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})\|)\}$, the S-estimator achieves an asymptotic breakdown point of $0 < r \leq 1/2$. We consider the common standard choices $r = 1/2$ and consistency for Σ at the normal model. Asymptotics can be found in Davies (1987), the influence function was calculated by Lopuhaä (1989) and efficiencies under normal distribution were calculated for instance in Croux and Haesbroeck (1999).

Table 3.4 lists the asymptotic relative efficiencies of the mentioned correlation estimators with respect to the Pearson correlation under normality. Specific tuning constants, parameters, weight functions, etc., are chosen as described above. The efficiency of the nonparametric estimators is declining with $|\rho|$, but the loss is rather small for moderate values. As we can see, the spatial correlation can well compete with highly robust estimators in terms of efficiency.

3.5.4 Simulations

We compare the correlation estimators in three different situations: under normality, under ellipticity and in outlier scenarios. We use the statistical software R (R Core Team, 2016), including the packages ICSNP (Nordhausen et al., 2012) (Tyler's M-estimator), MNM (Nordhausen and Oja, 2011) (elliptical power exponential distribution), mvtnorm Genz et al. (2014) (multivariate normal and elliptical t-distributions), pcaPP Filzmoser et al. (2011) (spatial median), rrcov Todorov and Filzmoser (2009) (S-estimator) and robustbase Rousseeuw et al. (2014)

	$\rho = 0$	$\rho = 0.5$		$\rho = 0$	$\rho = 0.5$
Pearson	1		GK-Q_n	0.823	
Spatial sign	0.5	0.464	GK-τ	0.8	
Gaussian rank	1		Tyler	0.5	
Spearman	0.912	0.867	rMCD	0.033	
Kendall	0.912	0.892	wMCD	0.401	
Quadrant	0.405	0.342	S	0.377	

Table 3.4: Asymptotic efficiency of correlation estimators for $p = 2$ under normality

(τ-scale, MCD and Q_n). In all scenarios, the estimators are transformed to be Fisher-consistent for the normal distribution. For some estimators, consistency-transformations for other distributions are known as well, but it is unrealistic in practice to know the kind of distribution in advance.

Throughout the simulations, we consider the one-stage spatial sign correlation (without pre-standardization) at distributions with equal marginal variances. The findings of Section 3.4 indicate that these simulation results may serve as a proxy for the two-stage estimator at elliptical distributions with arbitrary marginal variances.

Results under normality

First we examine bias and variance under normality. We choose $\rho = 0.5$, let the sample size n vary from 5 to 100, and generate 100,000 samples for each sample size. In Figure 3.6 (left), we see that all correlation estimators are biased towards zero in small samples. Next to the Pearson correlation, Kendall's τ and Spearman's ρ (the adequately transformed estimates) are least biased. The negative bias of the raw MCD still remains heavy even for $n = 100$. The spatial sign correlation behaves very well in terms of finite-sample variance. On the right-hand side of Figure 3.6, the variance times n (the n-stabilized variance) is plotted against n, which is, contrary to most other estimators, nearly a horizontal line. This indicates that asymptotic tests and confidence intervals based on the spatial sign correlation provide good approximations also in small samples.

Results under elliptical distributions

We consider two subclasses of elliptical distributions that generate varying tails: the t_ν-family and the power exponential family (e.g. Bilodeau and Brenner, 1999, pp. 208, 209). The results for the t_ν distribution are summarized in Table 3.5.4, where the mean squared error (MSE) of the various correlation estimators based on 100,000 samples are given for $\rho = 0$ and $\rho = 0.5$ and different degrees of freedom ν. Keep in mind that formally the correlation does not exist for one and two degrees of freedom, and we estimate the corresponding parameter ρ of the shape matrix instead, see also the beginning of Section 3.3. The MSE of the spatial sign correlation remains constant with respect to ν, which applies only to the quadrant correlation and Tyler's M-estimator among the other methods. For one degree of freedom and $\rho = 0.5$, the spatial sign correlation, together with Kendall's τ and Tyler's M-estimator, is most efficient. For one degree of freedom and $\rho = 0$, Spearman's ρ yields the smallest MSE by far. But this is due to its (asymptotic) bias towards zero. Contrary to Kendall's τ, the consistency transformation applied to Spearman's ρ under normality is not valid under ellipticity in general.

The MSEs, again based on 100,000 repetitions, for the power exponential distribution are displayed in Figure 3.7. The sample size is $n = 100$, the true correlation $\rho = 0.5$, and the power parameter α ranges from 0.02 to 2 in 56 (non-equidistant) steps. Letting $\alpha = 1$ corresponds to

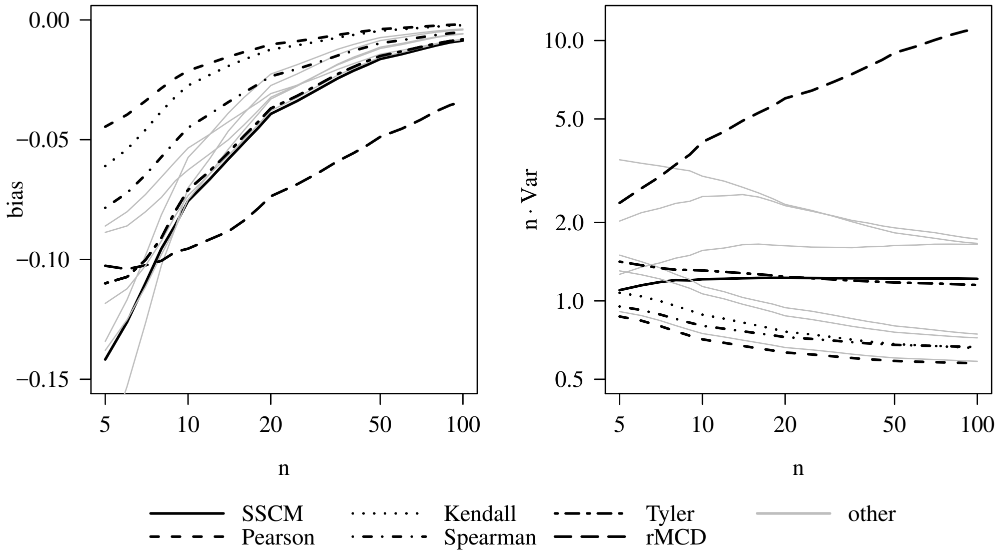

Figure 3.6: Simulated finite sample bias (left) and n·variance (right) under normality, $\rho = 0.5$ and different sample sizes n.

ρ	0				0.5			
ν	1	2	5	10	1	2	5	10
Pearson	0.356	0.112	0.021	0.013	0.283	0.077	0.012	0.007
Spatial sign	0.020	0.020	0.019	0.020	0.012	0.012	0.012	0.012
Quadrant	0.024	0.024	0.024	0.017	0.017	0.016	0.016	0.016
Kendall	0.019	0.016	0.013	0.012	0.012	0.010	0.008	0.007
Spearman	0.016	0.014	0.012	0.012	0.015	0.011	0.008	0.007
Gaussian rank	0.021	0.017	0.013	0.012	0.019	0.013	0.008	0.007
GK-Q_n	0.021	0.017	0.015	0.014	0.012	0.010	0.009	0.008
GK-τ	0.024	0.019	0.015	0.014	0.014	0.011	0.009	0.008
Tyler	0.020	0.020	0.020	0.020	0.012	0.012	0.012	0.012
rMCD	0.076	0.099	0.132	0.149	0.047	0.062	0.085	0.098
wMCD	0.037	0.035	0.034	0.032	0.022	0.021	0.020	0.019
S	0.033	0.030	0.029	0.029	0.019	0.017	0.017	0.017

Table 3.5: MSE under t_ν distributions with different degrees of freedom and $n = 100$.

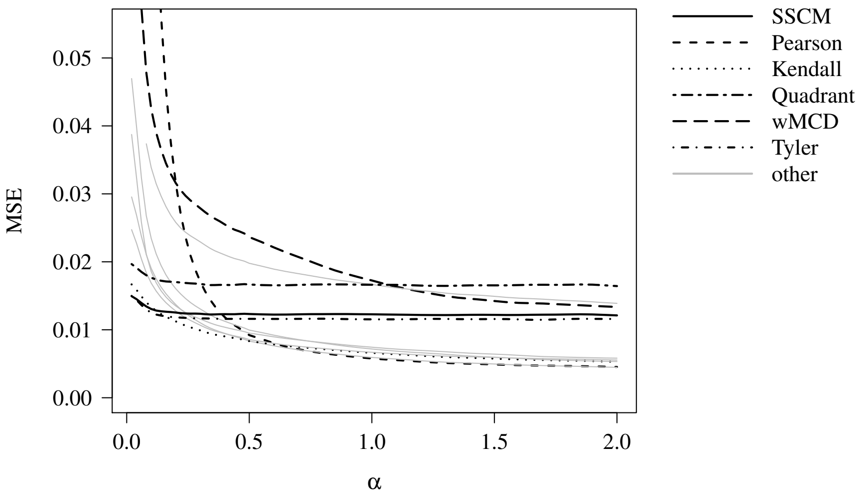

Figure 3.7: MSE of correlation estimators under the power exponential distribution with different α, $\rho = 0.5$ and $n = 100$.

the normal distribution and $\alpha = 0.5$ yields the elliptical Laplace distribution. With decreasing α, the distribution gets heavier tailed and more peaked in the origin, but all moments exist for any $\alpha > 0$ and the density remains bounded. As before, the MSE of the spatial correlation does not depend on the "tailedness parameter" α, which is in line with the asymptotic result. Only for very small α, we observe a slight incline. The power exponential distribution with a small α is particularly challenging for robust scatter estimators, since it possesses heavy tails and a probability mass concentration at the origin. Robust estimators downweight or reject outlying observations, which are in this case no contaminations, but carry the main information about the shape. In fact, the MSE of the raw MCD is above the displayed region in Figure 3.7. The spatial sign covariance matrix can cope well with such peaked distributions. For $\alpha < 0.1$, we find it, together with Tyler's estimator, to have the smallest MSE. However, it is crucial to use an appropriate location estimator that also works well with peaks at the center, see e.g. the discussion in Section 3.2 of Dürre et al. (2014). Altogether Kendall's τ appears to perform best over the whole range of α.

Results under contamination

To assess the robustness properties, we consider two scenarios: a single outlier of varying size, and an increasing amount of outliers stemming from a contamination distribution. In the first situation, we start from a bivariate normal sample with $\rho = 0.5$ and $n = 100$ and shift the first observation to the right by a distance h ranging from 0 to 5. This yields a high leverage point, suggesting a smaller correlation. We measure the influence of this one outlier in the x-direction by the difference of the estimate before and after the manipulation. The result is a sensitivity curve along the x-direction (divided by the factor $n = 100$), plotted in Figure 3.8 (left). The influence of the additive outlier is very small for the spatial sign correlation and also for most other robust estimators. An exception is the Gaussian rank correlation, which is known to have an unbounded influence function. Several highly robust estimators (in particular, the S-estimator and the MCD) completely disregard outliers that are sufficiently far away from the bulk of the data, and their sensitivity curves tend back to 0 as h further increases.

In the second setting, we start as usual with normally distributed data, $\rho = 0.5$, marginal variances 1 and $n = 100$. Then we replace, one after another, the "good" observations by outliers, which stem from a normal distribution with marginal variances 4 and correlation $\rho = -0.5$. On

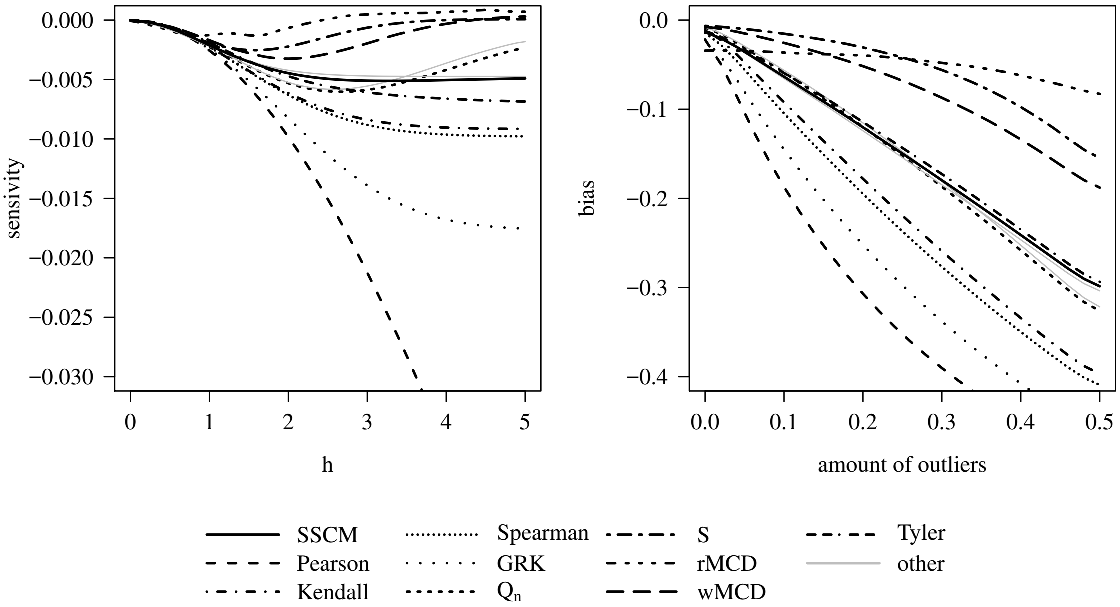

Figure 3.8: Bias of correlation estimators under normality with $\rho = 0.5$, $n = 100$. Left: one additive outlier of size h in the x-direction. Right: different fraction of outliers sampled from distribution with correlation $\rho = -0.5$.

the right-hand side of Figure 3.8, the bias of the estimators (average of 50,000 repetitions) is plotted against the contamination fraction. Here the picture is somewhat reversed to the efficiency results under normality: the rather efficient rank-based estimators like Spearman's ρ and Kendall's τ are substantially biased, and the rather inefficient and highly robust estimators (MCD, S) perform better. As before, the spatial sign correlation takes a place in the middle.

3.6 Multivariate spatial sign correlation matrix

One can construct an estimator of the correlation matrix R by filling the off-diagonal positions of the matrix estimate with the bivariate spatial sign correlation coefficients of all pairs of variables. Proposition 1 allows an alternative approach: First standardize the data marginally by a robust scale estimator and compute the SSCM of the transformed data. Then apply a singular value decomposition

$$\hat{S}_n(\boldsymbol{t}_n, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \hat{U}\hat{\Delta}\hat{U}^\top,$$

where $\hat{\Delta}$ contains the ordered eigenvalues $\hat{\delta}_1 \geq \ldots \geq \hat{\delta}_p$. One obtains estimates $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$ by inverting (3.8). Although theoretical results are yet to be established, we found in our simulations that the following fix point algorithm

$$\begin{aligned}
\hat{\lambda}_i^{(0)} &= \hat{\delta}_i, && i = 1, \ldots, p, \\
\tilde{\lambda}_i^{(k+1)} &= 2\hat{\delta}_i \left(\int_0^\infty \frac{1}{(1+\hat{\lambda}_i^{(k)}x)\prod_{j=1}^p (1+\hat{\lambda}_j^{(k)}x)^{\frac{1}{2}}} dx, \right)^{-1}, && i = 1, \ldots, p,\ k = 1, 2, \ldots \\
\hat{\lambda}_i^{(k+1)} &= \tilde{\lambda}_i^{(k+1)} \left(\sum_{j=1}^p \tilde{\lambda}_j^{(k+1)} \right)^{-1}, && i = 1, \ldots, p,\ k = 1, 2, \ldots
\end{aligned}$$

	n	100					1000				
	p	2	3	5	10	50	2	3	5	10	50
N	cor	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	sscor pairwise	1.9	1.9	1.9	1.9	1.9	2.0	2.0	2.0	2.0	2.0
	sscor multivariate	1.9	1.6	1.4	1.2	1.0	2.0	1.7	1.4	1.2	1.0
t_{10}	cor	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.4	1.3
	sscor pairwise	2.0	1.9	1.9	2.0	1.9	2.0	2.0	2.0	2.0	2.0
	sscor multivariate	2.0	1.7	1.3	1.2	1.0	2.0	1.7	1.4	1.2	1.0
t_5	cor	2.0	2.1	2.1	2.1	2.1	2.6	2.6	2.6	2.6	2.6
	sscor pairwise	2.0	2.0	1.9	2.0	1.9	2.1	2.0	2.0	2.0	2.0
	sscor multivariate	2.0	1.7	1.4	1.2	1.1	2.1	1.7	1.4	1.2	1.0
L	cor	1.6	1.5	1.3	1.2	1.1	1.6	1.5	1.3	1.2	1.1
	sscor pairwise	1.9	1.9	1.9	2.0	2.0	2.0	2.0	2.0	2.0	2.0
	sscor multivariate	1.9	1.6	1.4	1.2	1.1	2.0	1.7	1.4	1.2	1.1

Table 3.6: Simulated variances (multiplied by n) of one off-diagonal element of the correlation matrix estimate based on the moment correlation (cor), the pairwise spatial sign correlation (sscor pairwise) and the multivariate spatial sign correlation matrix (sscor multivariate) for spherical normal (N), t_5, t_{10}, and Laplace (L) distribution, several dimensions p and sample sizes $n = 100, 1000$.

works reliably and converges usually within 5 iterations if p is large. Let $\hat{\Lambda}$ denote the diagonal matrix containing $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$, then $\hat{V} = \hat{U}\hat{\Lambda}\hat{U}^T$ is a suitable estimator for the shape of the standardized data and $\hat{R}$ with $\hat{\rho}_{ij} = \hat{v}_{ij}/\sqrt{\hat{v}_{ii}\hat{v}_{jj}}$ an estimator for the correlation matrix, which we call the *multivariate spatial sign correlation matrix*. Contrary to the pairwise approach, the multivariate spatial sign correlation matrix is positive semi-definite by construction.

Theoretical properties of the new estimator are not straightforward to establish. By a small simulation study we want to obtain an impression of its efficiency. We compare the variances of the moment correlation, the pairwise as well as the multivariate spatial sign correlation under several elliptical distributions: normal, Laplace and t distributions with 5 and 10 degrees of freedom. The latter three generate heavier tails than the normal distribution. The Laplace distribution is obtained by the elliptical generator $g(x) = c_p \exp(-\sqrt{|x|}/2)$, where c_p is the appropriate integration constant depending on p (e.g. Bilodeau and Brenner, 1999, p. 209).

We take the identity matrix as shape matrix and compare the variances of an off-diagonal element of the matrix estimates for different dimensions $p = 2,\ 3,\ 5,\ 10,\ 50$ and sample sizes $n = 100,\ 1000$. We use the R packages mvtnorm (Genz et al., 2014) and MNM (Nordhausen and Oja, 2011) for the data generation. The results based on 10000 runs are summarized in Table 3.6.

Except for the moment correlation at the t_5 distribution, the results for $n = 100$ and $n = 1000$ are very similar. Note that the variance of the moment correlation decreases at the Laplace distribution as the dimension p increases, but not so for the other distributions considered. The lower dimensional marginals of the Laplace distribution are, contrary to the normal and the t-distributions, not Laplace distributed (see Kano, 1994), and the kurtosis of the one-dimensional marginals of the Laplace distribution in fact decreases as p increases.

Theorem 2 yields an asymptotic variance of 2 for the pairwise spatial sign correlation matrix elements regardless of the specific elliptical generator. This can also be observed in the simulation results. The moment correlation is twice as efficient under normality, but it has a higher variance at heavy tailed distributions. For uncorrelated t_5 distributed random variables, the spatial sign correlation outperforms the moment correlation. Looking at the multivariate spatial sign correlation, we see a strong increase of efficiency for larger p. For $p = 50$ the variance is comparable to that of the moment correlation. Since the asymptotic variance of the SSCM does not depend on the elliptical generator, this is expected to also hold for the multivariate spatial sign correlation, and this claim is confirmed by the simulations. The multivariate spatial sign correlation is more efficient than the moment correlation even under slightly heavier tails for moderately large p.

We simulate also from other shape matrices, e.g., the equi-correlation matrix

$$V = \begin{pmatrix} 1 & 0.5 & \dots & 0.5 \\ 0.5 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0.5 \\ 0.5 & \dots & 0.5 & 1 \end{pmatrix}.$$

The results can be found in Table 3.7. Except for the general smaller asymptotic variances we get the same picture. The asymptotic variance of the multivariate spatial sign correlation matrix is shrinking with growing dimension and approaches that of the sample correlation under normality, albeit more slowly than in the uncorrelated case.

	n	100					1000				
	p	2	3	5	10	50	2	3	5	10	50
N	cor	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6
	sscor pairwise	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2
	sscor multivariate	1.2	1.0	1.0	0.8	0.8	1.2	1.0	0.9	0.8	0.7
t_{10}	cor	0.8	0.7	0.7	0.8	0.8	0.8	0.7	0.8	0.7	0.8
	sscor pairwise	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2
	sscor multivariate	1.2	1.1	0.9	0.8	0.8	1.2	1.0	0.9	0.8	0.7
t_5	cor	1.2	1.2	1.2	1.2	1.2	1.5	1.5	1.5	1.5	1.5
	sscor pairwise	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2
	sscor multivariate	1.2	1.0	0.9	0.8	0.7	1.2	1.0	0.9	0.8	0.8
L	cor	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1
	sscor pairwise	1.2	1.2	1.2	1.3	1.2	1.2	1.2	1.2	1.2	1.2
	sscor multivariate	1.2	1.1	0.9	0.9	0.7	1.2	1.0	0.9	0.8	0.7

Table 3.7: Simulated variances (multiplied by n) of one off-diagonal element of the correlation matrix estimate based on the moment correlation (cor), the pairwise spatial sign correlation (sscor pairwise) and the multivariate spatial sign correlation matrix (sscor multivariate) for equi-correlated normal (N), t_5, t_{10}, and Laplace (L) distribution, several dimensions p and sample sizes $n = 100, 1000$.

An increase of efficiency for larger p is not uncommon for robust scatter estimators. It can be observed amongst others for M-estimators, the Tyler shape matrix, the MCD, and S-estimators (see e.g. Croux and Haesbroeck, 1999; Taskinen et al., 2006). All of these are affine equivariant estimators, requiring $n > p$. This is not necessary for the spatial sign correlation matrix.
One may expect that the efficiency gain for large p is at the expense of robustness. We therefore investigate the influence function of one off-diagonal element of the multivariate spatial sign correlation. The influence function is based on the concept that estimators are functionals working on distributions. In this setting the specific estimate based on a given dataset equals the functional evaluated at the corresponding empirical distribution. Denote by $\check{\rho}$ the functional representation of the multivariate spatial sign correlation with matrix-elements $\check{\rho}_{i,j}$, $1 \leq i <$

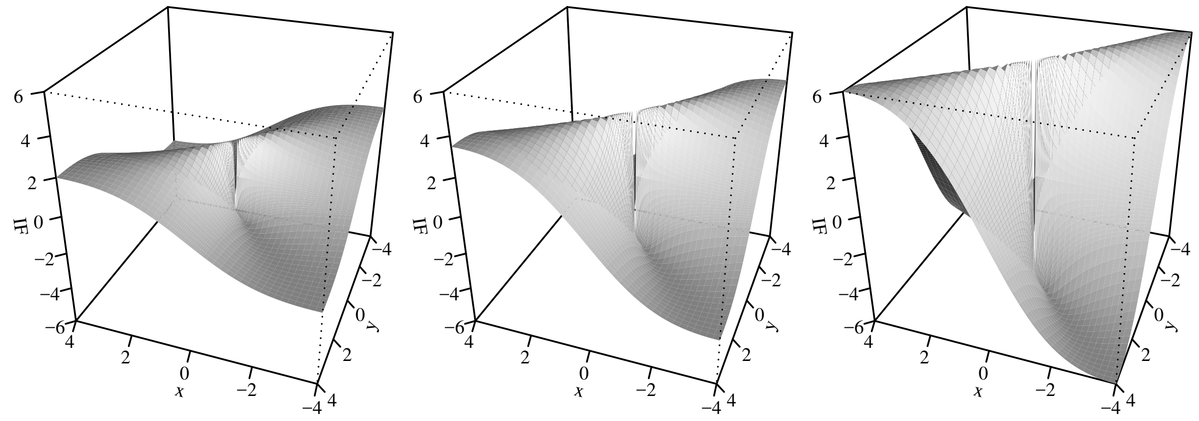

Figure 3.9: Partial influence functions of the off-diagonal element of multivariate spatial sign correlation $\breve{\rho}_{12}$ for $\mathbf{x} = (x, y, 0, \ldots, 0)$ under spherical distribution for $p = 2$ (left), $p = 5$ (center) and $p = 10$ (right).

$j \leq p$. Then the influence function $IF(\mathbf{x}, \breve{\rho}_{i,j}, F)$ is defined as

$$IF(\mathbf{x}, \breve{\rho}_{i,j}, F) = \lim_{\epsilon \to 0} \frac{\breve{\rho}_{i,j}((1-\epsilon)F + \epsilon\Delta_{\mathbf{x}}) - \breve{\rho}_{i,j}(F)}{\epsilon}$$

where $\Delta_{\mathbf{x}}$ denotes the Dirac measure putting its mass at $\mathbf{x}$. It measures the impact of an infinitesimal contamination at point $\mathbf{x}$ on $\breve{\rho}_{i,j}$ under distribution F. For further explanations and details about the influence function, see Hampel et al. (1986) and Huber and Ronchetti (2009).
Since we do not have an explicit representation for the estimated eigenvalues $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$, it seems to be challenging to calculate the influence function for arbitrary F and $\mathbf{x}$. Nevertheless, we can get results if we restrict ourselves to the case where F is elliptical with shape $V = I_p$ and $\mathbf{x}$ lies in a special hyperplane of $\mathbb{R}^p$. Furthermore we look at the case where the proportions of the marginal scales are known, respectively the data is not standardized prior to the computation of the SSCM.

Proposition 5. *(Dürre et al., 2017) Let* $F \in \mathscr{E}_2(\boldsymbol{t}, V)$ *and* $\breve{\rho}_{i,j}$ *denote the functional representation of the off-diagonal element of the multivariate spatial sign correlation without pre-standardization and let* $\mathbf{x} = (x, y, 0, \ldots, 0)^T$ *with* $x, y \in \mathbb{R}$, *then*

$$IF(\mathbf{x}, \breve{\rho}_{1,2}, F) = (p+2)\frac{xy}{x^2 + y^2}. \tag{3.18}$$

For $p = 2$, Proposition 5 is a special case of Proposition 4 which gives the influence function for arbitrary V. Although Proposition 5 is restricted to the situation where there is only contamination in the first two components, it provides evidence that the sensitivity of the multivariate spatial sign correlation increases with increasing dimension. One can see in Figure 3.9 respectively formula (3.18) that the influence functions are proportional to each other and that $|IF(\mathbf{x}, \breve{\rho}_{1,2}, F)|$ increases linearly in p for fixed $\mathbf{x} = (x, y, 0 \ldots, 0)$. This result indicates that the multivariate spatial sign correlation is more effected by outliers if p is large. Therefore the bivariate approach seems to be preferable if one suspects outliers and the dataset is high dimensional.

3.7 Conclusion

We have introduced a new bivariate correlation estimator based on the spatial sign covariance matrix, which we call the spatial sign correlation. We studied its asymptotic properties analytically and investigated its finite-sample and robustness properties by means of a comprehensive simulation study, comparing it to other popular robust correlation estimators. The simulations indicate that, although the spatial sign correlation at normality has a rather low relative efficiency of at most 1/2 with respect to the sample correlation, it is competitive when being compared to highly robust correlation measures that offer a similar resistance against outliers. The main advantage of the spatial sign correlation is its computational efficiency. It is therefore suited to be applied in a high-dimensional setting for pairwise correlation estimation. Robust pairwise correlation matrices have the advantage of a manageable computational effort as p increases, and that they work also for $n < p$. The price one usually has to pay, though, is the loss of the non-negative definiteness of the matrix estimate. For example, many nonparametric correlation matrix estimators (see Section 3.5) are based on an initial scatter matrix estimate which is non-negative definite, but not affine equivariant. The loss of non-negative definiteness occurs when a component-wise transformation is applied to render the entries consistent for the moment correlation. In applications where non-negative definiteness is important, one can "orthogonalize" the matrix estimate as suggested by Maronna and Zamar (2002), which involves an eigenvalue decomposition.

In the case of the spatial sign correlation, the pairwise approach means to compute bivariate spatial signs for every pair of observations. This entails the question if, when using the spatial median as location estimator, also bivariate spatial medians should be computed, or if rather each variable should be centered by the respective component of the p-variate spatial median. Computationally there is little obligation against the latter, the Weiszfeld algorithm (cf. e.g. Vardi and Zhang, 2001) scales nicely with p. From a statistical point of view, one might prefer the bivariate spatial median. Assume a trivariate elliptical model with uncorrelated margins and marginal variances 1, 1 and 100. The (trivariate) spatial median (as well as the SSCM) is very inefficient at such a strongly shaped distribution, and this inefficiency affects all of its components. In such a case, the bivariate spatial median is a more efficient location estimator for the first two components than the first two components of the trivariate spatial median. For a detailed and nice exposition of the efficiency properties of the spatial median see Magyar and Tyler (2011). Again, this effect can be reduced by pre-standardization.

The main drawback of the spatial sign correlation is the inefficiency under strongly shaped models, i.e., where the eigenvalues of the shape matrix strongly differ. The shapedness[4] due to different marginal scales may be eliminated by a componentwise standardization before computing the spatial sign correlation. We have shown that the resulting two-step estimator has the same asymptotic distribution as the spatial sign correlation applied to a sample from a model with equal marginal scales.

An important consequence is that the parameter a, the ratio of the marginal scales, drops from the expression for the asymptotic variance. The only parameter left is the generalized correlation coefficient ρ itself. This allows us to devise a variance-stabilizing transformation similar to Fisher's z-transformation. The confidence intervals obtained by this transform are very accurate also for very small samples. They are more precise in terms of coverage probabilities than those obtained by Fisher's z-transform and furthermore valid for all elliptical distributions. The prior standardization makes the spatial sign correlation really a practical estimator. The two-stage spatial sign correlation is implemented in the package sscor (Dürre and Vogel, 2016b)

[4] The term *shapedness* is used here for what is usually called *ellipticity*, i.e., the degree of divergence of an ellipse from a circle. We prefer to use *shapedness* since the term *ellipticity* in the context of elliptical distributions is often used to refer to the latter.

along with a correlation test including confidence intervals based on the h-transform.

The derivation of the spatial sign correlation hinges on the explicit connection between the spatial sign covariance matrix and the covariance matrix which is known under two-dimensional, elliptical distributions. The question naturally arises if and how these results may be extended to broader distributional assumption.

It is known that the covariance matrix and the shape matrix share the same eigenvectors not only under ellipticity but also within the broader class of linear transformations of permutation and sign-change invariant distributions, which includes the *symmetric independent-components model.* An elliptical vector X may be characterized by

$$\boldsymbol{X} = AR\boldsymbol{U} + \boldsymbol{\mu}, \tag{3.19}$$

where $A \in \mathbb{R}^{p\times p}$, $\boldsymbol{\mu} \in \mathbb{R}^p$, and R is a non-negative univariate random variable, independent of $\boldsymbol{U}$, which is uniformly distributed on the unit sphere in $\mathbb{R}^p$. The symmetric independent-components model may be similarly written as

$$\boldsymbol{X} = AY + \boldsymbol{\mu},$$

where A and $\boldsymbol{\mu}$ are as above, but $\boldsymbol{Y} = (Y^{(1)}, \ldots, Y^{(p)})^\top$ consists of i.i.d. symmetrically distributed components. However, an explicit connection between the eigenvalues of the SSCM and covariance matrix in this model is not known. Furthermore, spatial signs are not similarly invariant with respect to the distribution of $Y^{(1)}$ as they are with respect to the distribution of R in (3.19). A class of distributions to which this invariance extends is the *generalized elliptical family* as considered, e.g., by Frahm (2004). It is also generated by (3.19) with R being uniformly distributed on the unit sphere, but R and $\boldsymbol{U}$ need not be independent. However, the connection between the SSCM and the covariance matrix that holds under ellipticity does not extend to this class.

Next to the pairwise approach we propose a second estimator for the correlation matrix of an elliptical distributed random variable, the multivariate spatial sign correlation matrix. By a fixed-point algorithm one can invert the map between the eigenvalues of the shape and the spatial sign covariance matrix and, based on this, estimate the correlation matrix of an elliptical distributed random vector. We found the fixed-point algorithm to work reliably and fast for various shape matrices and dimensions. Simulations show that the resulting estimator is highly efficient in larger dimensions. Its asymptotic variance appears to approach that of the sample correlation under normality as the dimension is growing. Asymptotics confirming the simulation results are of great interest. The calculated partial influence function indicates that the efficiency gain of the spatial sign correlation matrix is at the cost of robustness. So the estimator does not seem to be very robust in the case of very high dimensions, but is nevertheless very efficient under heavy-tailed distributions.

3.8 Proofs

Proof of Proposition 1 and 2. We exercise the liberty to choose an appropriate distribution for Y and take the uniform distribution on the unit ball (not the unit sphere) with density $f(\boldsymbol{y}) = p/2\Gamma(p/2)\pi^{-p/2}\mathbb{1}_{[0,1]}(\boldsymbol{y}^\top\boldsymbol{y})$, which yields

$$\delta_i \;=\; \frac{p\Gamma(p/2)}{2\pi^{p/2}}\int_{\boldsymbol{y}^\top\boldsymbol{y}\le 1}\frac{\lambda_i y_i^2}{\lambda_1 y_1^2 + \ldots + \lambda_p y_p^2}d\boldsymbol{y} \;=\; \frac{2^p p\Gamma(p/2)}{2\pi^{p/2}}\int_{S_2}\frac{\lambda_i y_i^2}{\lambda_1 y_1^2 + \ldots + \lambda_p y_p^2}d\boldsymbol{y}$$

with $\boldsymbol{y} = (y_1, \ldots, y_p)$ and $S_{2,p} = \{\boldsymbol{y} \in \mathbb{R}^p \,|\, \boldsymbol{y}^\top\boldsymbol{y} \le 1, y_1, \ldots, y_p \ge 0\}$. Substituting $y_k = \sqrt{z_k}$, $1 \le k \le p$, we get

$$\delta_i = \frac{p\Gamma(p/2)}{2\pi^{p/2}}\int_{S_{1,p}}\frac{\lambda_i z_i}{\lambda_1 z_1 + \ldots + \lambda_p z_p}\prod_{j=1}^{p}\frac{1}{\sqrt{z_j}}dz \tag{3.20}$$

with $S_{1,p} = \{z \in \mathbb{R}^p \,|\, z_1 + \ldots + z_p \le 1, z \ge 0\}$. Now we apply formula 4.646 in Gradshteyn and Ryzhik (2000):

$$\int_{S_{1,n}} \frac{\prod_{k=1}^n x_k^{p_k-1}}{(\sum_{k=1}^n q_k x_k)^r} dx = \frac{\Gamma(p_1) \cdot \ldots \cdot \Gamma(p_n)}{\Gamma(\sum_{k=1}^n p_k - r + 1)\Gamma(r)} \int_0^\infty \frac{x^{r-1}}{\prod_{k=1}^n (1 + q_k x)^{p_k}} dx \tag{3.21}$$

for $p_1, \ldots, p_n, q_1, \ldots q_n > 0$, $p_1 + \ldots + p_n > r > 0$. Setting $n = p$, $r = 1$, $q_k = \lambda_k$ for $1 \le k \le p$, $p_i = 3/2$, and $p_k = 1/2$ for $k \ne i$, we obtain from (3.20) the expression for δ_i given in Proposition 1. As for η_{ii}, we proceed similarly. Choosing again the uniform density on the unit ball and substituting $y_k = \sqrt{z_k}$, $1 \le k \le p$, yields

$$\eta_{ii} = \frac{p\Gamma(p/2)}{2\pi^{p/2}} \int_{S_{1,p}} \frac{\lambda_i^2 z_i^2}{(\lambda_1 z_1 + \ldots + \lambda_p z_p)^2} \prod_{j=1}^p \frac{1}{\sqrt{z_j}} dz.$$

Applying (3.21) with $n = p$, $r = 2$, $q_k = \lambda_k$ for $1 \le k \le p$, $p_i = 5/2$, and $p_k = 1/2$ for $k \ne i$, we obtain the expression for η_{ii} as given in Proposition 2. As for η_{ij} with $i \ne j$, we obtain a similar expression, to which we apply (3.21) with the same parameters except $p_i = p_j = 3/2$, and $p_k = 1/2$ for $k \ne i, j$. This completes the proof. □

Towards the proof of Proposition 3, we consider as an intermediate step the SSCM-based shape estimator $\hat{V}_n$ defined at the beginning of Section 3.3. Precisely, we give the asymptotic distribution of the estimator

$$\hat{V}_{0,n} = \begin{pmatrix} \hat{v}_{0,11} & \hat{v}_{0,12} \\ \hat{v}_{0,12} & \hat{v}_{0,22} \end{pmatrix} = \frac{1}{\sqrt{\hat{v}_{11}\hat{v}_{22}}} \hat{V}_n.$$

We have remarked at the end of Section 3.1 that, for analyzing the scale-invariant properties of the shape of an elliptical distribution, fixing the overall scale of the shape matrix V is not necessary, and we view the shape as an equivalence class of positive definite matrices being proportional to each other. For explicit computations, however, it is at some point necessary to fix the scale, that is, picking one specific representative from the equivalence class. Various ways of standardizing the shape can be found in the literature. Paindaveine (2008) argues to choose $\det(V) = 1$, which corresponds to our choice of $\hat{V}_n$ in Section 3.3. However, for our purposes, it is most convenient to standardize V such that the product of its diagonal elements is 1, which corresponds to $\hat{V}_{0,n}$ described above. Accordingly, we denote by V_0 the representative of the equivalence class with this property and parametrize it as

$$V_0 = \begin{pmatrix} a & \rho \\ \rho & a^{-1} \end{pmatrix}, \tag{3.22}$$

where the parameters a and ρ have the same meaning as in Section 3.3, that is, the ratio of the diagonal elements of V and the correlation, respectively. Denote $\hat{s}_{ij}$, $i, j = 1, 2$ the elements of the SSCM $\hat{S}(\mathbb{X}_n; \hat{\boldsymbol{\mu}}_n)$ and let $d = \frac{1}{2} + \sqrt{(\hat{s}_{11} - 1/2)^2 + \hat{s}_{12}^2}$. An eigenvalue decomposition yields

$$\hat{\Delta}_n = \begin{pmatrix} d & 0 \\ 0 & 1-d \end{pmatrix} \quad \text{and} \quad \hat{U}_n = \begin{pmatrix} \frac{\hat{s}_{12}}{\sqrt{\hat{s}_{12}^2 + (d - \hat{s}_{11})^2}} & \frac{\hat{s}_{12}}{\sqrt{\hat{s}_{12}^2 + (1 - d - \hat{s}_{11})^2}} \\ \frac{d - \hat{s}_{11}}{\sqrt{\hat{s}_{12}^2 + (d - \hat{s}_{11})^2}} & \frac{1 - d - \hat{s}_{11}}{\sqrt{\hat{s}_{12}^2 + (1 - d - \hat{s}_{11})^2}} \end{pmatrix}$$

and by choosing $\hat{\lambda}_1 = \hat{\delta}_1/\hat{\delta}_2$ as well as $\hat{\lambda}_2 = \hat{\delta}_2/\hat{\delta}_1$ one arrives at $\hat{v}_{0,11} = \frac{\hat{U}_{n_{11}}^2 \hat{\lambda}_1 + \hat{U}_{n_{21}}^2 \hat{\lambda}_2}{\hat{U}_{n_{12}}^2 \hat{\lambda}_1 + \hat{U}_{n_{22}}^2 \hat{\lambda}_2}$ where the nominator simplifies to

$$\begin{aligned} \hat{U}_{n_{11}}^2 \hat{\lambda}_1 + \hat{U}_{n_{21}}^2 \hat{\lambda}_2 &= \frac{\hat{U}_{n_{11}}^2 d^2 + \hat{U}_{n_{21}}^2 (1-d)^2}{(1-d)d} \\ &= \frac{d^2 + d\hat{U}_{n_{11}}^2 + (1-d)\hat{U}_{n_{21}}^2 - d(\hat{U}_{n_{11}}^2 + \hat{U}_{n_{21}}^2)}{d(1-d)} = \frac{\hat{s}_{11}}{d(1-d)} - 1 \end{aligned}$$

and the denominator to $\hat{U}_{n_{12}}^2 \hat{\lambda}_1 + \hat{U}_{n_{22}}^2 \hat{\lambda}_2 = \frac{1-\hat{s}_{11}}{d(1-d)} - 1$ such that

$$\hat{v}_{0,11} = \frac{\hat{s}_{11}^2 + \hat{s}_{12}^2}{\hat{s}_{12}^2 + (1-\hat{s}_{11})^2}. \tag{3.23}$$

The nominator of $\hat{v}_{0,12} = \frac{\hat{U}_{n_{11}}\hat{\lambda}_1\hat{U}_{n_{21}} + \hat{U}_{n_{12}}\hat{\lambda}_2\hat{U}_{n_{22}}}{\sqrt{\hat{U}_{n_{12}}^2\hat{\lambda}_1 + \hat{U}_{n_{22}}^2\hat{\lambda}_2}\sqrt{\hat{U}_{n_{11}}^2\hat{\lambda}_1 - \hat{U}_{n_{21}}^2\hat{\lambda}_2}}$ equals

$$\begin{aligned}\hat{U}_{n_{11}}\hat{\lambda}_1\hat{U}_{n_{21}} + \hat{U}_{n_{12}}\hat{\lambda}_2\hat{U}_{n_{22}} &= \frac{\hat{U}_{n_{11}} d \hat{U}_{n_{21}} + \hat{U}_{n_{12}}(1-d)\hat{U}_{n_{22}}}{(1-d)d} - (\hat{U}_{n_{11}}\hat{U}_{n_{21}} + \hat{U}_{n_{12}}\hat{U}_{n_{22}}) \\ &= \frac{\hat{s}_{12}}{d(1-d)},\end{aligned}$$

hence

$$\hat{v}_{0,12} = \frac{\hat{s}_{12}}{\sqrt{(\hat{s}_{12}^2 + \hat{s}_{11}^2)(\hat{s}_{12}^2 + (1-\hat{s}_{11})^2)}}. \tag{3.24}$$

The following proposition summarizes the asymptotic behavior of the estimator $\hat{V}_{0,n}$.

Proposition A1. *Under the assumptions of Proposition 3, we have for $n \to \infty$ that*

(1) $\hat{V}_{0,n} \xrightarrow{a.s.} V_0$ *and*

(2) $\sqrt{n}\left\{(\hat{v}_{0,11}, \hat{v}_{0,12})^T - (a, \rho)^T\right\} \xrightarrow{d} N_2\left(\mathbf{0}, W_{V_0}\right)$, *where* $W_{V_0} = G W_S G^T$ *with*

$$G = \begin{pmatrix} 2a + \left(a^2+1\right)\sqrt{1-\rho^2} & 0 & \left(1-a^2\right)\rho & 0 \\ a^{-1}\left(a^2-1\right)\rho\sqrt{1-\rho^2} & 0 & \sqrt{1-\rho^2}\{2 + (a^2+1)a^{-1}\sqrt{1-\rho^2}\} & 0 \end{pmatrix}, \tag{3.25}$$

and W_S is defined by (3.10).

Proof of Proposition A1. Part (1) is a consequence of the continuous mapping theorem, part (2) follows with the delta method. Note that V_0 is specified by the two elements $\hat{v}_{0,11}$ and $\hat{v}_{0,12}$, and, likewise, $\hat{S}_n = \hat{S}_n(\mathbb{X}_n; \boldsymbol{\mu}_n)$ by the two elements $\hat{s}_{11}$ and $\hat{s}_{12}$. Let H be the function that maps $(\hat{s}_{11}, \hat{s}_{12})$ to $(\hat{v}_{0,11}, \hat{v}_{0,12})$ and (s_{11}, s_{12}) to (a, ρ). It is given explicitly by the formulas (3.23) and (3.24), from which we can compute its derivative. However, it turns out that it is easier to compute the derivative of its inverse and apply the inverse function theorem. With $\{(s_{11}, s_{12}) \,|\, 0 < s_{11} < 1, |s_{12}| < \sqrt{s_{11}(1-s_{11})}\}$ and $\{(a, \rho) \,|\, 0 < a < \infty, |\rho| < 1\}$ being its domain and image, respectively, the function H is invertible and continuously differentiable. Let J denote its inverse. The function J maps (a, ρ) to (s_{11}, s_{12}) and is described by (3.2) and (3.6). In the following, we will compute its derivate, for which we require an explicit form of J. The eigenvalue decomposition of V_0 can be computed by the computer algebra system Maxima (2014) and is given by $\lambda_{1/2} = (2a)^{-1}\left(a^2 + 1 \pm \sqrt{q}\right)$ and

$$U = \begin{pmatrix} \frac{\sqrt{\sqrt{q}+a^2-1}}{\sqrt{2}q^{\frac{1}{4}}} & \frac{\sqrt{\sqrt{q}-a^2+1}}{\sqrt{2}g^{\frac{1}{4}}} \\ \frac{s(\rho)\sqrt{\sqrt{q}-a^2+1}}{\sqrt{2}q^{\frac{1}{4}}} & -\frac{s(\rho)(\sqrt{\sqrt{q}+a^2-1}}{\sqrt{2}q^{\frac{1}{4}}} \end{pmatrix}, \tag{3.26}$$

where $q = 4a^2\rho^2 + (a^2-1)^2$ and $\boldsymbol{s}(\cdot)$ denotes the sign function defined at the beginning of Section 3.2. By (3.2) and (3.6) we find

$$\begin{aligned} s_{11} &= U_{11}^2 \frac{\sqrt{\lambda_1}}{\sqrt{\lambda_1}+\sqrt{\lambda_2}} + U_{12}^2 \frac{\sqrt{\lambda_2}}{\sqrt{\lambda_1}+\sqrt{\lambda_2}} \\ &= \sqrt{k}\frac{\sqrt{q}+a^2-1}{2\sqrt{q}(\sqrt{m}+\sqrt{k})} + \sqrt{m}\frac{\sqrt{q}-a^2+1}{2\sqrt{q}(\sqrt{m}+\sqrt{k})} \\ &= \frac{\sqrt{k}\{4a^2\rho^2+\sqrt{q}(a^2-1)+(a^2-1)^2\} + \sqrt{m}\{4a^2\rho^2+\sqrt{q}(1-a^2)+(a^2-1)^2)\}}{2q(\sqrt{m}+\sqrt{k})} \end{aligned}$$

and

$$\begin{aligned} s_{12} &= U_{1,1}U_{2,1}\frac{\sqrt{\lambda_1}}{\sqrt{\lambda_1}+\sqrt{\lambda_2}} + U_{1,2}U_{2,2}\frac{\sqrt{\lambda_2}}{\sqrt{\lambda_1}+\sqrt{\lambda_2}} \\ &= \frac{s(\rho)\sqrt{q-(a^2-1)^2}}{2\sqrt{q}}\frac{\sqrt{k}}{\sqrt{k}+\sqrt{m}} - \frac{s(\rho)\sqrt{q-(a^2-1)^2}}{2\sqrt{q}}\frac{\sqrt{m}}{\sqrt{k}+\sqrt{m}} \\ &= s(\rho)\frac{2a|\rho|}{2\sqrt{q}}\frac{(\sqrt{k}-\sqrt{m})^2}{k-m} \\ &= (2q)^{-1}a\rho(\sqrt{k}-\sqrt{m})^2, \end{aligned}$$

where $k = a^2+1+\sqrt{q}$ and $m = a^2+1-\sqrt{q}$. The derivative of J can also be computed by Maxima and is

$$\mathbb{D}J(a,\rho) = \begin{pmatrix} \frac{2a(a^2+1)\rho^2\sqrt{1-\rho^2}+(a^2-1)^2(1-\rho^2)}{q((a^2+1)\sqrt{1-\rho^2}+2a(1-\rho^2))} & -\frac{(a-1)a(a+1)\rho(2a\sqrt{1-\rho^2}-a^2-1)}{q((a^2+1)\sqrt{1-\rho^2}+2a(1-\rho^2))} \\ -\frac{(a-1)(a+1)\rho((a^2+1)\sqrt{1-\rho^2}-2a(1-\rho^2))}{q((a^2+1)\sqrt{1-\rho^2}+2a(1-\rho^2))} & \frac{a((a^2-1)^2\sqrt{1-\rho^2}+2a(a^2+1)\rho^2)}{q((a^2+1)\sqrt{1-\rho^2}+2a(1-\rho^2))} \end{pmatrix}.$$

Straightforward calculation yields that the determinant of this matrix equals $\det \mathbb{D}J(a,\rho) = a\sqrt{1-\rho^2}\{\left(a^2+1\right)\,\sqrt{1-\rho^2}+2\,a\,\left(1-\rho^2\right)\}^{-2}$. By virtue of the inverse function theorem, we have $\mathbb{D}H(s_{11},s_{12}) = (\mathbb{D}J(a,\rho))^{-1}$. Hence we obtain $\mathbb{D}H(s_{11},s_{12})$ by inverting the 2×2 matrix $\mathbb{D}J(a,\rho)$. It can be seen to be (except for the zero columns) the matrix G in Proposition A1. The proof is complete. □

Proof of Proposition 3. Proposition 3 is an immediate corollary of Proposition A1, noting that $\hat{\rho}_n = \hat{v}_{0,12}$. By (3.10) and (3.26) we get

$$W_S = \frac{\frac{1}{2}a\sqrt{1-\rho^2}(a^2+1)+a^2(\rho^2-1)}{4a^2\rho^2+(a^2-1)^2}\begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix}.$$

The asymptotic variance of $\hat{\rho}_n$ is the then lower diagonal element of $W_{V_0} = GW_SG^T$

$$\begin{aligned} W_{V_0 2,2} &= \frac{\frac{1}{2}a\sqrt{1-\rho^2}(a^2+1)+a^2(\rho^2-1)}{4a^2\rho^2+(a^2-1)^2}(G_{2,1}^2+G_{2,3}^2) \\ &= \frac{(1-\rho^2)^{3/2}a}{4a^2\rho^2+(a^2-1)^2}\left[\frac{(a^2+1)^3}{2a^2} - 4a(1-\rho^2)^{3/2} - 2(1-\rho^2)(a^2+1) + \frac{(a^2+1)^2\sqrt{1-\rho^2}}{a}\right] \\ &= (1-\rho^2)^2 + (1-\rho^2)^{3/2}(a^2+1)/2/a. \end{aligned}$$

□

Proof of Proposition 4. Croux et al. (2010) give the influence function of the off-diagonal element of the SSCM for $p = 2$. Calculation of the diagonal elements is straightforward, and we obtain for $F \in \mathscr{E}_2(\boldsymbol{\mu}, V)$:

$$IF(\boldsymbol{x}, \hat{S}, F) = \boldsymbol{x}\boldsymbol{x}^T/(\boldsymbol{x}^T\boldsymbol{x}) - S(F). \tag{3.27}$$

Applying the chain rule and using the derivatives calculated in the proof of Proposition A1 above, we arrive at the influence function of the spatial sign correlation as given in Proposition 4. □

In the proof of Theorem 1, we make use of the following lemma, which states that the difference between the empirical versions of Γ_j with and without location estimation converges to zero in probability.

Lemma A1. *Let $\boldsymbol{t} \in \mathbb{R}^p$ and $\boldsymbol{X}$ be a p-variate random vector with distribution F satisfying*

(I) $\mathrm{E}|\boldsymbol{X} - \boldsymbol{t}|^{-2/3} < \infty$.

Let further $\mathbb{X}_n = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^\top$ be an iid sample drawn from F and $\boldsymbol{t}_n$ a series of p-variate random vectors satisfying

(II) $\sqrt{n}(\boldsymbol{t}_n - \boldsymbol{t}) = O_P(1)$.

Finally, let A be a diagonal $p \times p$ matrix with positive diagonal entries. Then, for all $1 \le j \le p$,

$$\frac{1}{n}\sum_{i=1}^n \left[a_j^2(X_i^{(j)} - t_n^{(j)})^2 \frac{A(\boldsymbol{X}_i - \boldsymbol{t}_n)(\boldsymbol{X}_i - \boldsymbol{t}_n)^\top A}{\{(\boldsymbol{X}_i - \boldsymbol{t}_n)^\top A^2 (\boldsymbol{X}_i - \boldsymbol{t}_n)\}^2}\right] - \frac{1}{n}\sum_{i=1}^n \left[a_j^2(X_i^{(j)} - t^{(j)})^2 \frac{A(\boldsymbol{X}_i - \boldsymbol{t})(\boldsymbol{X}_i - \boldsymbol{t})^\top A}{\{(\boldsymbol{X}_i - \boldsymbol{t})^\top A^2 (\boldsymbol{X}_i - \boldsymbol{t})\}^2}\right]$$

converges to zero in probability as $n \to \infty$.

Proof. To shorten notation and without loss of generality we will assume that $\boldsymbol{t} = 0$ and $A = I_p$. We will show componentwise convergence, i.e.

$$\frac{1}{n}\sum_{i=1}^n \left\{ \frac{(X_i^{(j)} - t_n^{(j)})^2 (\boldsymbol{X}_i - \boldsymbol{t}_n)^{(k)} (\boldsymbol{X}_i - \boldsymbol{t}_n)^{(l)}}{|\boldsymbol{X}_i - \boldsymbol{t}_n|^4} - \frac{(X_i^{(j)})^2 X_i^{(k)} X_i^{(l)}}{|\boldsymbol{X}_i|^4} \right\} \xrightarrow{p} 0 \tag{3.28}$$

as $n \to \infty$ for all $1 \le j, k, l \le p$. We use the following random partition of $\mathbb{R}^p$:

$$B_n = \{\boldsymbol{x} \in \mathbb{R}^p | \; |\boldsymbol{x} - \boldsymbol{t}_n| \ge \frac{1}{2}|\boldsymbol{x}|\}, \qquad B_n^C = \{\boldsymbol{x} \in \mathbb{R}^p | \; |\boldsymbol{x} - \boldsymbol{t}_n| < \frac{1}{2}|\boldsymbol{x}|\}. \tag{3.29}$$

and the corresponding random partition of the index set $\{1, \ldots, n\}$:

$$I_n = \{1 \le i \le n | \boldsymbol{X}_i \in B_n\}, \qquad I_n^C = \{1 \le i \le n | \boldsymbol{X}_i \in B_n^C\}.$$

Letting K_i denote the summands in (3.28), we write

$$\frac{1}{n}\sum_{i=1}^n K_i = \frac{1}{n}\sum_{i \in I_n} K_i + \frac{1}{n}\sum_{i \in I_n^C} K_i. \tag{3.30}$$

For the second sum on the right-hand side of (3.30) we make use of $|K_i| \le 2$ to obtain $|n^{-1}\sum_{i \in I_n^C} K_i| \le 2n^{-1}\sum_{i=1}^n \mathbb{1}_{B_n^C}(\boldsymbol{X}_i)$. The right-hand side of the last inequality is shown to converge to zero in probability under the assumptions of Lemma A1 as in the proof of Theorem 1 in Dürre et al. (2014). The first sum on the right-hand side of (3.30) is decomposed into

$$\frac{1}{n}\sum_{i \in I_n} K_i \;=\; \frac{1}{n}\sum_{i \in I_n} \frac{\{(X_i^{(j)} - t_n^{(j)})^2 - (X_i^{(j)})^2\}(\boldsymbol{X}_i - \boldsymbol{t}_n)^{(k)}(\boldsymbol{X}_i - \boldsymbol{t}_n)^{(l)}|\boldsymbol{X}_i|^4}{|\boldsymbol{X}_i - \boldsymbol{t}_n|^4 |\boldsymbol{X}_i|^4}$$

$$+\ \frac{1}{n}\sum_{i\in I_n}\frac{(X_i^{(j)})^2 t_n^{(k)}(\boldsymbol{X}_i-\boldsymbol{t}_n)^{(l)}|\boldsymbol{X}_i|^4}{|\boldsymbol{X}_i-\boldsymbol{t}_n|^4|\boldsymbol{X}_i|^4}\ +\ \frac{1}{n}\sum_{i\in I_n}\frac{(X_i^{(j)})^2 X_i^{(k)} t_n^{(l)}|\boldsymbol{X}_i|^4}{|\boldsymbol{X}_i-\boldsymbol{t}_n|^4|\boldsymbol{X}_i|^4}$$

$$+\ \frac{1}{n}\sum_{i\in I_n}\frac{(X_i^{(j)})^2 X_i^{(k)} X_i^{(l)}(|\boldsymbol{X}_i|^4-|\boldsymbol{X}_i-\boldsymbol{t}_n|^4)}{|\boldsymbol{X}_i-\boldsymbol{t}_n|^4|\boldsymbol{X}_i|^4}.$$

Call the four terms from left to right $\mathcal{T}_1$, $\mathcal{T}_2$, $\mathcal{T}_3$, $\mathcal{T}_4$. Since $X_i \in B_n$ implies $|X_i| \leq 2|X_i - t_n|$, we have

$$\begin{aligned}
|\mathcal{T}_1| &\leq \frac{1}{n}\sum_{i\in I_n}\left|\frac{t_n^{(j)}(t_n^{(j)}-2X_i^{(j)})(\boldsymbol{X}_i-\boldsymbol{t}_n)^{(k)}(\boldsymbol{X}_i-\boldsymbol{t}_n)^{(l)}|\boldsymbol{X}_i|^4}{|\boldsymbol{X}_i-\boldsymbol{t}_n|^4|\boldsymbol{X}_i|^4}\right| \\
&\leq \frac{1}{n}\sum_{i\in I_n}\left|\frac{t_n^{(j)}(t_n^{(j)}-X_i^{(j)})}{|\boldsymbol{X}_i-\boldsymbol{t}_n|^2}\right| + \frac{1}{n}\sum_{i\in I_n}\left|\frac{t_n^{(j)}X_i^{(j)}}{|\boldsymbol{X}_i-\boldsymbol{t}_n|^2}\right| \\
&\leq \frac{2}{n}\sum_{i\in I_n}\frac{|t_n^{(j)}|}{|\boldsymbol{X}_i|} + \frac{4}{n}\sum_{i\in I_n}\frac{|t_n^{(j)}|}{|\boldsymbol{X}_i|}\ \leq\ \frac{6}{n}\sum_{i=1}^{n}\frac{|t_n^{(j)}|}{|\boldsymbol{X}_i|}\ =\ 6\sqrt{n}|t_n^{(j)}|\left\{\frac{1}{n^{3/2}}\sum_{i=1}^{n}\frac{1}{|\boldsymbol{X}_i|}\right\}\xrightarrow{p}0,
\end{aligned}$$

since the term in $\{\cdot\}$ converges to zero almost surely by Marczinkiewicz's law of large numbers (Loève, 1977, p. 255). Convergence to zero of the remaining terms $\mathcal{T}_2$, $\mathcal{T}_3$ and $\mathcal{T}_4$ is shown analogously. The proof of Lemma A1 is complete. □

Remark: One can see from the last displayed line that, similarly to Theorem 1 of Dürre et al. (2014), the lemma can be proven also under slightly different conditions. For instance, assumption (II) can weakened to $\boldsymbol{t}_n \xrightarrow{p} \boldsymbol{t}$ in exchange for the stronger moment condition $\mathrm{E}|\boldsymbol{X}-\boldsymbol{t}|^{-1} < \infty$.

We are now ready to prove Theorem 1.

Proof of Theorem 1. Let $\tilde{\boldsymbol{t}}_n(\mathbb{X}_n) = A_n^{-1}\boldsymbol{t}_n(\mathbb{X}_nA_n)$[5] and write $\sqrt{n}\{\hat{S}_n(\mathbb{X}_nA_n,\boldsymbol{t}_n(\cdot))-\hat{S}_n(\mathbb{X}_nA,A\boldsymbol{t})\}$ as

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{\{A_n\boldsymbol{X}_i-A_n\tilde{\boldsymbol{t}}_n(\mathbb{X}_n)\}\{A_n\boldsymbol{X}_i-A_n\tilde{\boldsymbol{t}}_n(\mathbb{X}_n)\}^\top}{\{A_n\boldsymbol{X}_i-A_n\tilde{\boldsymbol{t}}_n(\mathbb{X}_n)\}^\top\{A_n\boldsymbol{X}_i-A_n\tilde{\boldsymbol{t}}_n(\mathbb{X}_n)\}}-\frac{\{A\boldsymbol{X}_i-A\tilde{\boldsymbol{t}}_n(\mathbb{X}_n)\}\{A\boldsymbol{X}_i-A\tilde{\boldsymbol{t}}_n(\mathbb{X}_n)\}^\top}{\{A\boldsymbol{X}_i-A\tilde{\boldsymbol{t}}_n(\mathbb{X}_n)\}^\top\{A\boldsymbol{X}_i-A\tilde{\boldsymbol{t}}_n(\mathbb{X}_n)\}}\right]$$

$$+\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{\{A\boldsymbol{X}_i-A\tilde{\boldsymbol{t}}_n(\mathbb{X}_n)\}\{A\boldsymbol{X}_i-A\tilde{\boldsymbol{t}}_n(\mathbb{X}_n)\}^\top}{\{A\boldsymbol{X}_i-A\tilde{\boldsymbol{t}}_n(\mathbb{X}_n)\}^\top\{A\boldsymbol{X}_i-A\tilde{\boldsymbol{t}}_n(\mathbb{X}_n)\}}-\frac{\{A\boldsymbol{X}_i-A\boldsymbol{t}\}\{A\boldsymbol{X}_i-A\boldsymbol{t}\}^\top}{\{A\boldsymbol{X}_i-A\boldsymbol{t}\}^\top\{A\boldsymbol{X}_i-A\boldsymbol{t}\}}\right].$$

Call the first term $\mathcal{T}_1$ and the second $\mathcal{T}_2$. The convergence of $\mathcal{T}_2$ to zero in probability follows with Dürre et al. (2014, Theorem 2). Let $\tilde{\boldsymbol{X}}_i = A\boldsymbol{X}_i$, $\boldsymbol{\tau}_n = A\tilde{\boldsymbol{t}}_n$ and $\boldsymbol{\tau} = A\boldsymbol{t}$. Then Theorem 2 of Dürre et al. (2014) essentially states that $\sqrt{n}\{\hat{S}_n(\tilde{\mathbb{X}}_n,\boldsymbol{\tau}_n)-\hat{S}_n(\tilde{\mathbb{X}}_n,\boldsymbol{\tau})\}\xrightarrow{p}0$, where $\tilde{\mathbb{X}}_n = (\tilde{\boldsymbol{X}}_1,\ldots,\tilde{\boldsymbol{X}}_n)^\top$. This is not stated explicitly in the text of the theorem, but this is what is proven. To check that the assumptions are met, note that by Conditions (C3), (C4) and (C5) we have

$$\begin{aligned}
&\sqrt{n}(\boldsymbol{\tau}_n-\boldsymbol{\tau}) = A\sqrt{n}\{A_n^{-1}\boldsymbol{t}_n(\mathbb{X}_nA_n)-t\} \\
&=\ A\sqrt{n}\{A_n^{-1}\boldsymbol{t}_n(\mathbb{X}_nA_n)-\boldsymbol{t}_n(\mathbb{X}_n)\}\ +\ A\sqrt{n}\{\boldsymbol{t}_n(\mathbb{X}_n)-\boldsymbol{t}\} \\
&=\ \sqrt{n}AA_n^{-1}\{\boldsymbol{t}_n(\mathbb{X}_nA_n)-A_n\boldsymbol{t}_n(\mathbb{X}_n)\}\ +\ A\sqrt{n}\{\boldsymbol{t}_n(\mathbb{X}_n)-\boldsymbol{t}\}\ =\ O_P(1).
\end{aligned}$$

[5] Technically, $\tilde{\boldsymbol{t}}_n$ is a function of $\mathbb{X}_n$ as well as A_n. We can understand $\tilde{\boldsymbol{t}}_n(\mathbb{X}_n)$ as a short-hand notation, where the dependence on A_n is simply suppressed, but the notation is also justified in the sense that A_n usually is a function of $\mathbb{X}_n$.

The latter is sufficient (along with continuity of F), see the remarks below Theorem 3 in Dürre et al. (2014). We are thus left to prove $\mathcal{T}_1 \xrightarrow{d} \Xi_p$. Let $\boldsymbol{Y}_i = \boldsymbol{X}_i - \tilde{\boldsymbol{t}}_n(\mathbb{X}_n)$, where we suppress the dependence the on n in this short-hand notation. Then $\mathcal{T}_1$ can be further decomposed into

$$\mathcal{T}_1 \;=\; \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{A_n\boldsymbol{Y}_i\boldsymbol{Y}_i^\top A_n - A\boldsymbol{Y}_i\boldsymbol{Y}_i^\top A}{\boldsymbol{Y}_i^\top A^2\boldsymbol{Y}_i} \;+\; \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\{\boldsymbol{Y}_i^\top(A^2-A_n^2)\boldsymbol{Y}_i\}\,A_n\boldsymbol{Y}_i\boldsymbol{Y}_i^\top A_n}{\boldsymbol{Y}_i^\top A_n^2\boldsymbol{Y}_i\boldsymbol{Y}_i^\top A^2\boldsymbol{Y}_i}.$$

We call the terms $\mathcal{T}_{1,a}$ and $\mathcal{T}_{1,b}$, where we have

$$\mathcal{T}_{1,a} \;=\; A_nA^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{A\boldsymbol{Y}_i\boldsymbol{Y}_i^\top A}{\boldsymbol{Y}_i^\top A^2\boldsymbol{Y}_i}\right)A^{-1}\sqrt{n}(A_n-A)+\sqrt{n}(A_n-A)A^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{A\boldsymbol{Y}_i\boldsymbol{Y}_i^\top A}{\boldsymbol{Y}_i^\top A^2\boldsymbol{Y}_i}\right),$$

which converges in distribution to $S(F_0,0)A^{-1}Z + ZA^{-1}S(F_0,0)$, since

$$\frac{1}{n}\sum_{i=1}^{n}\frac{A\boldsymbol{Y}_i\boldsymbol{Y}_i^\top A}{\boldsymbol{Y}_i^\top A^2\boldsymbol{Y}_i} \xrightarrow{p} S(F_0,0)$$

by Theorem 1 in Dürre et al. (2014). Writing $\mathcal{T}_{1,b}$ as $\mathcal{T}_{1,b} = \mathcal{L} + \mathcal{R}$ with

$$\mathcal{L} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{2\{\boldsymbol{Y}_i^\top(A-A_n)A\}\,\boldsymbol{Y}_iA\boldsymbol{Y}_i\boldsymbol{Y}_i^\top A}{(\boldsymbol{Y}_i^\top A^2\boldsymbol{Y}_i)^2},$$

$$\mathcal{R} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{\{\boldsymbol{Y}_i^\top(A-A_n)(A+A_n)\boldsymbol{Y}_i\}\,A_n\boldsymbol{Y}_i\boldsymbol{Y}_i^\top A_n}{\boldsymbol{Y}_i^\top A_n^2\boldsymbol{Y}_i\boldsymbol{Y}_i^\top A^2\boldsymbol{Y}_i} - 2\frac{\{\boldsymbol{Y}_i^\top(A-A_n)A\boldsymbol{Y}_i\}\,A\boldsymbol{Y}_i\boldsymbol{Y}_i^\top A}{(\boldsymbol{Y}_i^\top A^2\boldsymbol{Y}_i)^2}\right],$$

we find for $\mathcal{L}$ by using Lemma A1

$$\mathcal{L} = 2\sum_{j=1}^{p}\{A^{-1}\sqrt{n}(A-A_n)\}^{(j,j)}\frac{1}{n}\sum_{i=1}^{n}\{(AY_i)^{(j)}\}^2\frac{A\boldsymbol{Y}_i\boldsymbol{Y}_i^\top A}{(\boldsymbol{Y}_i^\top A^2\boldsymbol{Y}_i)^2} \xrightarrow{d} -2\sum_{j=1}^{p}(A^{-1}Z)^{(j,j)}\Gamma_j.$$

It remains to show that $\mathcal{R}$ vanishes asymptotically. We further decompose $\mathcal{R}$ into

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{\{\boldsymbol{Y}_i^\top(A-A_n)(A+A_n)\boldsymbol{Y}_i\}\,A_n\boldsymbol{Y}_i\boldsymbol{Y}_i^\top A_n}{\boldsymbol{Y}_i^\top A_n^2\boldsymbol{Y}_i\boldsymbol{Y}_i^\top A^2\boldsymbol{Y}_i} - \frac{\{\boldsymbol{Y}_i^\top(A-A_n)(A+A_n)\boldsymbol{Y}_i\}\,A_n\boldsymbol{Y}_i\boldsymbol{Y}_i^\top A_n}{\boldsymbol{Y}_i^\top A^2\boldsymbol{Y}_i\boldsymbol{Y}_i^\top A^2\boldsymbol{Y}_i}\right]$$

$$+\;\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{\{\boldsymbol{Y}_i^\top(A-A_n)(A+A_n)\boldsymbol{Y}_i\}\,A_n\boldsymbol{Y}_i\boldsymbol{Y}_i^\top A_n}{\boldsymbol{Y}_i^\top A^2\boldsymbol{Y}_i\boldsymbol{Y}_i^\top A^2\boldsymbol{Y}_i} - \frac{\{\boldsymbol{Y}_i^\top(A-A_n)2A\boldsymbol{Y}_i\}\,A_n\boldsymbol{Y}_i\boldsymbol{Y}_i^\top A_n}{\boldsymbol{Y}_i^\top A^2\boldsymbol{Y}_i\boldsymbol{Y}_i^\top A^2\boldsymbol{Y}_i}\right]$$

$$+\;\frac{1}{\sqrt{n}}\sum_{i=1}^{n}2\left[\frac{\{\boldsymbol{Y}_i^\top(A-A_n)A\boldsymbol{Y}_i\}\,A_n\boldsymbol{Y}_i\boldsymbol{Y}_i^\top A_n}{\boldsymbol{Y}_i^\top A^2\boldsymbol{Y}_i\boldsymbol{Y}_i^\top A^2\boldsymbol{Y}_i} - \frac{\{\boldsymbol{Y}_i^\top(A-A_n)A\boldsymbol{Y}_i\}\,A\boldsymbol{Y}_i\boldsymbol{Y}_i^\top A_n}{\boldsymbol{Y}_i^\top A^2\boldsymbol{Y}_i\boldsymbol{Y}_i^\top A^2\boldsymbol{Y}_i}\right]$$

$$+\;\frac{1}{\sqrt{n}}\sum_{i=1}^{n}2\left[\frac{\{\boldsymbol{Y}_i^\top(A-A_n)A\boldsymbol{Y}_i\}\,A\boldsymbol{Y}_i\boldsymbol{Y}_i^\top A_n}{\boldsymbol{Y}_i^\top A^2\boldsymbol{Y}_i\boldsymbol{Y}_i^\top A^2\boldsymbol{Y}_i} - \frac{\{\boldsymbol{Y}_i^\top(A-A_n)A\boldsymbol{Y}_i\}\,A\boldsymbol{Y}_i\boldsymbol{Y}_i^\top A}{\boldsymbol{Y}_i^\top A^2\boldsymbol{Y}_i\boldsymbol{Y}_i^\top A^2\boldsymbol{Y}_i}\right]$$

and denote the four terms by $\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_3$ and $\mathcal{S}_4$, respectively. For $\mathcal{S}_1$ we get

$$|\mathcal{S}_1| \le \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left|\frac{\{\boldsymbol{Y}_i^\top(A-A_n)(A+A_n)\boldsymbol{Y}_i\}^2A_n\boldsymbol{Y}_i\boldsymbol{Y}_i^\top A_n}{\boldsymbol{Y}_i^\top A_n^2\boldsymbol{Y}_i(\boldsymbol{Y}_i^\top A^2\boldsymbol{Y}_i)^2}\right|$$

$$\le \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{\boldsymbol{Y}_i^\top(A-A_n)(A+A_n)\boldsymbol{Y}_i}{\boldsymbol{Y}_i^\top A^2\boldsymbol{Y}_i}\right\}^2$$

$$= \frac{1}{\sqrt{n}} \sum_{j=1}^{p} \sum_{k=1}^{p} \{\sqrt{n}(A - A_n)(A + A_n)\}^{(j,j)} \{\sqrt{n}(A - A_n)(A + A_n)\}^{(k,k)} \frac{1}{n} \sum_{i=1}^{n} \left(\frac{Y_i^{(j)} Y_i^{(k)}}{\boldsymbol{Y}_i^\top A^2 \boldsymbol{Y}_i} \right)^2,$$

which converges to zero in probability. For $\mathcal{S}_2$, we obtain

$$\begin{aligned}
\mathcal{S}_2 &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\{\boldsymbol{Y}_i^\top (A - A_n)(A_n - A)\boldsymbol{Y}_i\}\, A_n \boldsymbol{Y}_i \boldsymbol{Y}_i^\top A_n}{\boldsymbol{Y}_i^\top A^2 \boldsymbol{Y}_i \boldsymbol{Y}_i^\top A^2 \boldsymbol{Y}_i} \\
&= \frac{1}{\sqrt{n}} \sum_{j=1}^{p} -\{a_j^{-1} \sqrt{n}(A_n - A)^{(j,j)}\}^2\, A_n A^{-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} (a_j Y_i^{(j)})^2 \frac{A \boldsymbol{Y}_i \boldsymbol{Y}_i^\top A}{(\boldsymbol{Y}_i^\top A^2 \boldsymbol{Y}_i)^2} \right\} A^{-1} A_n \\
&\xrightarrow{d} 0 \cdot \sum_{j=1}^{p} -(Z^{(j,j)}/a_j)^2 \Gamma_j,
\end{aligned}$$

where we have again used Lemma A1. Similar calculations yield that $\mathcal{S}_3 = o_P(1)$ and $\mathcal{S}_4 = o_P(1)$ as $n \to \infty$. Note that, although we have treated $\mathcal{T}_{1,a}$ and $\mathcal{L}$ individually, they converge in fact jointly. Both are essentially linear functions of $\sqrt{n}(A_n - A)$. The proof of Theorem 1 is complete. □

Proof of Corollary 1. As in Theorem 1, let $\boldsymbol{X}_0 = A(\boldsymbol{X} - \boldsymbol{t})$. Then $\boldsymbol{X}_0 \sim F_0 \in \mathscr{E}_2(0, V_0)$. Since V_0 has equal diagonal elements, its eigenvalue decomposition is given by $V_0 = U \Lambda U^\top$, where

$$U = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}, \qquad \Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} = c \begin{pmatrix} 1 - \rho & 0 \\ 0 & 1 + \rho \end{pmatrix}. \tag{3.31}$$

for some $c > 0$. Hence, by (3.2) and (3.6), we have

$$S(F_0, 0) = \begin{pmatrix} 1/2 & \delta \\ \delta & 1/2 \end{pmatrix}$$

with $\delta = (1 - \sqrt{1 - \rho^2})/(2\rho)$ if $\rho \neq 0$ and $\delta = 0$ otherwise, and hence

$$A^{-1} Z S(F_0, 0) + S(F_0, 0) A^{-1} Z = \begin{pmatrix} Z_1/a_1 & (Z_1/a_1 + Z_2/a_2)\delta \\ (Z_1/a_1 + Z_2/a_2)\delta & Z_2/a_2 \end{pmatrix}. \tag{3.32}$$

To compute the remaining part $-2\sum_{j=1}^{2} (Z_j/a_j)\Gamma_j$, we have to evaluate the integrals Γ_j, $j = 1, 2$. Towards this end, we write $\boldsymbol{X}_0 = U \Lambda^{1/2} \boldsymbol{Y}$, where U and Λ are as in (3.31) and $\boldsymbol{Y}$ has a spherical distribution, and consider the matrix

$$\begin{aligned}
W &= \mathrm{E} \left\{ \mathrm{vec} \left(\frac{\boldsymbol{X}_0 \boldsymbol{X}_0^\top}{\boldsymbol{X}_0^\top \boldsymbol{X}_0} \right) \mathrm{vec} \left(\frac{\boldsymbol{X}_0 \boldsymbol{X}_0^\top}{\boldsymbol{X}_0^\top \boldsymbol{X}_0} \right)^\top \right\} \\
&= (U \otimes U) \mathrm{E} \left\{ \mathrm{vec} \left(\frac{\Lambda^{1/2} \boldsymbol{Y} \boldsymbol{Y}^\top \Lambda^{1/2}}{\boldsymbol{Y}^\top \Lambda \boldsymbol{Y}} \right) \mathrm{vec} \left(\frac{\Lambda^{1/2} \boldsymbol{Y} \boldsymbol{Y}^\top \Lambda^{1/2}}{\boldsymbol{Y}^\top \Lambda \boldsymbol{Y}} \right)^\top \right\} (U \otimes U)^\top
\end{aligned}$$

The expectation on the right-hand side is independent of the elliptical generator g and is given as an explicit function of λ_1 and λ_2 in the proof of Proposition 2(3) in Dürre et al. (2015b). Plugging in our specific forms of Λ and U, see (3.31), we obtain

$$W = \begin{pmatrix} \alpha & \beta & \beta & \gamma \\ \beta & \gamma & \gamma & \beta \\ \beta & \gamma & \gamma & \beta \\ \gamma & \beta & \beta & \alpha \end{pmatrix}$$

with

$$\alpha = \frac{\sqrt{1-\rho^2}+2\rho^2-1}{4\rho^2}, \quad \beta = \frac{1-\sqrt{1-\rho^2}}{4\rho} = \delta/2, \quad \gamma = \frac{1-\sqrt{1-\rho^2}}{4\rho^2}$$

if $\rho \neq 0$, and $\alpha = 3/8$, $\beta = 0$, $\gamma = 1/8$ if $\rho = 0$. Since W contains Γ_1 as upper diagonal block and Γ_2 as lower diagonal block, we obtain

$$-2\left(\frac{Z_1}{a_1}\Gamma_1 + \frac{Z_2}{a_2}\Gamma_2\right) = -2\begin{pmatrix} \frac{Z_1}{a_1}\alpha + \frac{Z_2}{a_2}\gamma & \left(\frac{Z_1}{a_1} + \frac{Z_2}{a_2}\right)\beta \\ \left(\frac{Z_1}{a_1} + \frac{Z_2}{a_2}\right)\beta & \frac{Z_1}{a_1}\gamma + \frac{Z_2}{a_2}\alpha \end{pmatrix}. \tag{3.33}$$

Putting (3.32) and (3.33) together, we finally arrive at

$$\Xi_2 = \begin{pmatrix} Z_1/a_1 - Z_2/a_2 & 0 \\ 0 & Z_2/a_2 - Z_1/a_1 \end{pmatrix}$$

which completes the proof of Corollary 1. □

For the proof of Theorem 2 we require a slight generalization of the delta method.

Lemma A2. *Let $(\boldsymbol{U}_n)_{n\in\mathbb{N}}$ be a series of p-dimensional random vectors and $(a_n)_{n\in\mathbb{N}}$ a sequence of real numbers such that $a_n \to \infty$ as $n \to \infty$ and*

(I) $a_n(\boldsymbol{U}_n - \boldsymbol{u}) = O_p(1)$ as $n \to \infty$ for some $\boldsymbol{u} \in \mathbb{R}^p$. Let furthermore

(II) $h : \mathbb{R}^p \to \mathbb{R}$ be continuously differentiable at $\boldsymbol{u} = (u_1, \dots, u_p)^\top$ with $\frac{\partial h(\boldsymbol{u})}{\partial u_i} = 0$ for all $i \in I$ for some subset $I \subset \{1, \dots, p\}$, and

(III) $a_n[\boldsymbol{U}_n - \boldsymbol{u}]_{I^C} \xrightarrow{d} \boldsymbol{\Psi}$, where $[\boldsymbol{U}_n - \boldsymbol{u}]_{I^C}$ denotes the random vector obtained from $\boldsymbol{U}_n - \boldsymbol{u}$ by deleting all components in I.

Then $a_n\{h(\boldsymbol{U}_n) - h(\boldsymbol{u})\} \xrightarrow{d} [h'(\boldsymbol{u})]_{I^C}\boldsymbol{\Psi}$.

If $I = \emptyset$, Lemma A2 boils down to the usual delta method. If some components of $h'(\boldsymbol{u})$ are zero (which are gathered in the index set I), it suffices to ensure the joint convergence of the remaining components of $a_n(\boldsymbol{U}_n - \boldsymbol{u})$ and the boundedness in probability of $a_n(\boldsymbol{U}_n - \boldsymbol{u})$ to conclude the convergence of $a_n\{h(\boldsymbol{U}_n) - h(\boldsymbol{u})\}$.

Proof of Lemma A2. The proof is similar to the proof of Lemma 5.3.2. in Bickel and Doksum (2001). Since h is continuously differentiable, for every $\epsilon > 0$ there exists a $\delta > 0$ such that

$$|\boldsymbol{u} - \boldsymbol{v}| \le \delta \;\Rightarrow\; |h(\boldsymbol{v}) - h(\boldsymbol{u}) - h'(\boldsymbol{u})(\boldsymbol{v} - \boldsymbol{u})| \le \epsilon|\boldsymbol{v} - \boldsymbol{u}|. \tag{3.34}$$

Condition (I) implies that $\boldsymbol{U}_n \xrightarrow{p} \boldsymbol{u}$, i.e., $P(|\boldsymbol{U}_n - \boldsymbol{u}| \le \delta) \to 1$. Thus using (3.34), we have for every $\epsilon > 0$ that $P(|h(\boldsymbol{U}_n) - h(\boldsymbol{u}) - h'(\boldsymbol{u})(\boldsymbol{U}_n - \boldsymbol{u})| \le \epsilon|\boldsymbol{U}_n - \boldsymbol{u}|) \to 1$ which implies $a_n\{h(\boldsymbol{U}_n) - h(\boldsymbol{u}) - h'(\boldsymbol{u})(\boldsymbol{U}_n - \boldsymbol{u})\} = o_p(|a_n(\boldsymbol{U}_n - \boldsymbol{u})|) = o_p(1)$. The latter may be re-written as

$$a_n\{h(\boldsymbol{U}_n) - h(\boldsymbol{u})\} = a_n h'(\boldsymbol{u})(\boldsymbol{U}_n - \boldsymbol{u}) + o_p(1),$$

and the result follows by Conditions (II) and (III) and Slutsky's lemma. □

Proof of Theorem 2. We write

$$\sqrt{n}(\hat{\rho}_{\sigma,n} - \rho) = \sqrt{n}\left[\gamma\{\operatorname{vec}\hat{S}_n(\mathbb{X}_n A_n, \boldsymbol{t}_n(\cdot))\} - \gamma\{\operatorname{vec} S(F_0, 0)\}\right],$$

where F_0 is, as in Theorem 1, the distribution of $\boldsymbol{X}_0 = A(\boldsymbol{X} - \boldsymbol{t})$, and $\gamma : \mathbb{R}^4 \to \mathbb{R}$ is the function that maps the (vectorized) two-dimensional spatial sign covariance matrix of an elliptical distribution to the corresponding generalized correlation coefficient. The function γ is given by (3.11). Its derivative γ' equals the second row of G in 3.25 evaluated at $a = 1$ since F_0 has equal marginal scales

$$\gamma'\{\operatorname{vec} S(F_0, 0)\} = \begin{pmatrix} 0 & 0 & 2\sqrt{1-\rho^2}(1+\sqrt{1-\rho^2}) & 0 \end{pmatrix}.$$

We further decompose

$$\begin{aligned} &\sqrt{n}\operatorname{vec}\left\{\hat{S}_n(\mathbb{X}_n A_n, \boldsymbol{t}_n(\cdot)) - S(F_0, 0)\right\} \qquad (3.35) \\ = \;& \sqrt{n}\operatorname{vec}\left\{\hat{S}_n(\mathbb{X}_n A_n, \boldsymbol{t}_n(\cdot)) - \hat{S}_n(\mathbb{X}_n A, A\boldsymbol{t})\right\} \;+\; \sqrt{n}\operatorname{vec}\left\{\hat{S}_n(\mathbb{X}_n A, A\boldsymbol{t}) - S(F_0, 0)\right\}, \end{aligned}$$

where we call the two terms on the right hand side $\mathcal{T}_1$ and $\mathcal{T}_2$. We deduce two things: First,

$$\sqrt{n}\operatorname{vec}\left\{\hat{S}_n(\mathbb{X}_n A_n, \boldsymbol{t}_n(\cdot)) - S(F_0, 0)\right\} = O_p(1) \qquad \text{as } n \to \infty,$$

since $\mathcal{T}_1 \xrightarrow{d} \Xi_2$ by Theorem 1, and $\mathcal{T}_2$ converges in distribution as a corollary of the central limit theorem (or as a special case of Proposition 2 in Dürre et al. (2015b)). Second, the third component of (3.35) converges in distribution to the same limit as $\mathcal{T}_2^{(3)}$, since $\mathcal{T}_1^{(3)}$ converges to zero in probability by Corollary 1. Here we use $(\,\cdot\,)^{(3)}$ to denote the third component of a vector. The asymptotic distribution of $\mathcal{T}_2$ is given by Proposition 2 in Dürre et al. (2015b). Making use of the particular structure of V_0, i.e., equal diagonal elements, see (3.31), we obtain $\mathcal{T}_2^{(3)} \xrightarrow{d} \mathcal{N}(0, w)$ with $w = (\sqrt{1-\rho^2} + \rho^2 - 1)/(2\rho)^2$ if $\rho \neq 0$ and $w = 1/8$ if $\rho = 0$. Applying Lemma A2 with γ in the role of h, and $I^C = \{3\}$, we obtain

$$\sqrt{n}(\hat{\rho}_{\sigma,n} - \rho) \xrightarrow{d} [\gamma'\{\operatorname{vec} S(F_0, 0)\}]_{(1,3)} \cdot \mathcal{N}(0, w) = \mathcal{N}(0, (1-\rho^2)^2 + (1-\rho^2)^{3/2}).$$

since

$$\begin{aligned} \frac{\sqrt{1-\rho^2} + \rho^2 - 1}{4\rho^2} 4(1-\rho^2)(1+\sqrt{1-\rho^2})^2 =& \frac{1-(1-\rho^2)}{\rho^2}(1-\rho^2)^{\frac{3}{2}}(1+\sqrt{1-\rho^2}) \\ =& (1-\rho^2)^2 + (1-\rho^2)^{\frac{3}{2}}. \end{aligned}$$

Note also that $\gamma'(\cdot)$ is a 1×4 matrix. The proof of Theorem 2 is complete. □

Proof of Corollary 2. By the delta method, the function h has to satisfy

$$|h'(x)| = \{(1-x^2)^2 + (1-x^2)^{3/2}\}^{-1/2}. \qquad (3.36)$$

The function h given in Corollary 2 fulfills this requirement and is further strictly increasing and odd. To find the antiderivative of (3.36), we have used the compute algebra system Maxima (2014). Substituting $z = 1 - \sqrt{1-x^2}$ yields the integral $\int\{\sqrt{(1-z)z}(2-z)\}^{-1}dz$, for which Maxima gives the primitive $2^{-1/2}\arcsin\{(3z-2)/|z-2|\}$. □

Proof of Proposition 5: Let denote $\check{S}$ the functional representation of the spatial sign covariance matrix $\check{S}(F) = \mathbb{E}_F\left(\frac{\boldsymbol{X}\boldsymbol{X}^\top}{\boldsymbol{X}^\top\boldsymbol{X}}\right)$ where $\boldsymbol{X}$ has distribution F. Since $\check{S}((1-\epsilon)F + \epsilon\Delta_{\boldsymbol{x}}) = (1-\epsilon)I_p + \epsilon\boldsymbol{x}\boldsymbol{x}^\top$ is a block diagonal matrix, we get the following eigenvalue decomposition $\check{S}((1-\epsilon)F + \epsilon\boldsymbol{x}\boldsymbol{x}^\top) = U_{xy}\Delta_\epsilon U_{xy}^\top$ where

$$U_{x,y} = \begin{pmatrix} \frac{x}{\sqrt{x^2+y^2}} & \frac{y}{\sqrt{x^2+y^2}} & 0 & \dots & 0 \\ \frac{y}{\sqrt{x^2+y^2}} & \frac{-x}{\sqrt{x^2+y^2}} & 0 & \dots & 0 \\ 0 & 0 & 1 & & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & & 1 \end{pmatrix} \quad \text{and} \quad \Delta_\epsilon = \begin{pmatrix} \frac{1+(p-1)\epsilon}{p} & 0 & 0 & \dots & 0 \\ 0 & \frac{1-\epsilon}{p} & 0 & \dots & 0 \\ 0 & 0 & \frac{1-\epsilon}{p} & & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & & \frac{1-\epsilon}{p} \end{pmatrix}.$$

We need to know how the perturbation of the eigenvalues of the SSCM translates into the eigenvalues of the shape matrix. The function $\Phi : \mathbb{R}^p \to \mathbb{R}^p$ which maps the eigenvalues of the shape to the eigenvalues of the SSCM is injective (see Proposition 1 in Dürre et al., 2016). Therefore the shape matrix related to Δ_ϵ contains only two distinct eigenvalues: λ_1 and $\lambda_2 = \ldots = \lambda_p$. We can simplify the situation even further since the eigenvalues are not uniquely defined and standardize them such that $\lambda_2 = \ldots, \lambda_p = 1$. On the other hand we have $\sum_{i=1}^p \delta_i = 1$ and therefore $\delta_i = \frac{1-\delta_1}{p-1}$, $i = 2, \ldots, p$. Consequently in this case the connection between the eigenvalues can be expressed by the one-dimensional function $f : [0,1] \to [0,\infty)$ which maps the first eigenvalue of Δ_ϵ to the first of the shape matrix.

Let $\zeta : \mathbb{R}^{p\times p} \to [-1,1]$ denote the function which computes the correlation coefficient between the first and second component given the shape matrix: $\zeta(A) = \frac{a_{12}}{\sqrt{a_{11}a_{22}}}$ and denote further $k(\epsilon) = \frac{1+(p-1)\epsilon}{p}$, then straightforward calculations yields,

$$\begin{aligned} &\lim_{\epsilon\to 0} \frac{\check{\rho}_{i,j}((1-\epsilon)F + \epsilon\Delta_{\boldsymbol{x}}) - \check{\rho}_{i,j}(F)}{\epsilon} \\ &= \lim_{\epsilon\to 0} \frac{\zeta\left[U_{xy}\begin{pmatrix} f\left(\frac{1+(p-1)\epsilon}{p}\right) & & 0 \\ & 1 & \\ & & \ddots & \\ 0 & & & 1 \end{pmatrix}U_{xy}^\top\right] - \zeta\left[U_{xy}I_pU_{xy}^\top\right]}{\epsilon} \\ &= \lim_{\epsilon\to 0} \frac{1}{\epsilon} \frac{(f[k(\epsilon)]-1)xy}{\sqrt{y^2 + f[k(\epsilon)]x^2}\sqrt{x^2 + f[k(\epsilon)]y^2}} =: \frac{\partial}{\partial\epsilon} h(f[k(\epsilon)])\bigg|_{\epsilon=0}. \end{aligned}$$

By the chain rule we get:

$$\frac{\partial}{\partial\epsilon} h(f[k(\epsilon)])\bigg|_{\epsilon=0} = \frac{\partial}{\partial\epsilon} h(x)\bigg|_{x=1} \cdot \frac{\partial}{\partial y} f(y)\bigg|_{y=1/p} \cdot \frac{\partial}{\partial\epsilon} k(\epsilon)\bigg|_{\epsilon=0}.$$

Whereas differentiation of h and k is straightforward, we do not have an explicit representation of f. Since we only need its derivative, we can apply the inverse function theorem. Using (3.8) and Leibniz's rule we arrive at

$$\begin{aligned} \frac{\partial}{\partial x} f(x)\bigg|_{x=1/p} &= \frac{1}{\frac{\partial}{\partial x} f^{-1}(x)|_{x=1}} \\ &= 1/\left(\frac{1}{2}\int_0^\infty \frac{1}{(1+z)^{\frac{p}{2}+1}} dz - \frac{3}{4}\int_0^\infty \frac{z}{(1+z)^{\frac{p}{2}+2}} dz\right) =: \frac{1}{A_1 + A_2}. \end{aligned}$$

For A_1 and A_2 we can apply formula 3.193-3 in Gradshteyn and Ryzhik (2000):

$$\int_0^\infty \frac{x^{\mu-1}dx}{(1+\beta x)^\nu}\, dx = B(\mu, \nu - \mu) \quad \text{for} \quad \nu > \mu > 0$$

where B denotes the beta function. Setting $\beta = 1,\ \mu = 1$ and $\nu = p/2 + 1$ for A_1 respectively $\mu = 2$ and $\nu = p/2 + 2$ for A_2 and using the relationship between beta and gamma function we arrive at $A_1 = \frac{1}{p}$ and $A_2 = \frac{3}{2p(p/2+1)}$. Straightforward term manipulations yield the stated formula (3.18). □

Chapter 4

Robust change-point detection in panel data

4.1 Introduction

There is an increasing amount of literature on panel data taking structural breaks into account, see for example Im et al. (2005), Bai and Carrion-I-Silvestre (2009), Karavias and Tzavalis (2012) and Baltagi et al. (2016). Often the work of Joseph and Wolfson (1992) is regarded as starting point of change point detection in panel data. Therein two different change point models are introduced, namely the common change point model, where the time of change $\theta_i \in \{1, \ldots, T\}$ is identical for every individual $i = 1, \ldots, N$, and the random change point model, where θ_i is independent and identically distributed following an unknown distribution P_θ. Recent work mostly considers the first approach. Joseph and Wolfson (1992), De Wachter and Tzavalis (2005), De Wachter and Tzavalis (2012) and Baltagi et al. (2017) consider homogeneous panels, where either the dependence structure or the noise distribution is the same for all individuals, whereas Bai (2010) and Kim (2011) look also at heterogeneous ones.

Surprisingly little attention has been paid to robust change-point procedures. To the best of our knowledge, the only exception is the article of Joseph and Wolfson (1992) where a robust variation of their original test procedure, a kind of Mann-Whitney-Wilcoxon test, is proposed. The lack of robust methods is in contrast to change point detection in one-dimensional (see for example Csörgo and Horváth, 1987; Hušková, 1996; Dehling et al., 2015) or multi-dimensional time series (see Koziol, 1978; Quessy et al., 2013; Vogel and Fried, 2015). In these settings robust procedures are not only more reliable in case of some corrupted observations, but they also turn out to be more powerful under heavy tailed distributions (see Dehling et al., 2017; Dehling et al., 2015).

In this chapter we propose a robust test for the fundamental problem of a common change point in location. As opposed to Joseph and Wolfson (1992) our test can cope with heterogeneous panels and short range dependence. Like Bai (2010), Horváth and Hušková (2012) and Jirak et al. (2015) we consider the case where both the time dimension T and the cross-sectional dimension N tend to infinity. In contrast to them, by choosing a bounded ψ-function, moment assumptions are not required. Our test is based on M-estimation and can be seen as a generalization of the test proposed in Horváth and Hušková (2012).

This chapter is structured as follows: in the next section we define the test statistic and explain how to choose the required tuning parameters. We present theoretical properties and conditions in Section 4.3. Section 4.4 contains a small simulation study showing the usefulness of the procedure and its finite sample performance. All proofs are deferred to Section 4.6.

4.2 Testing procedure

Let $(X_{i,t})_{i\in\{1,\ldots,N\},\ t\in\{1,\ldots,T\}}$ denote a panel where N is the number of individuals, which are observed at T equidistant time points. We assume the simple structure

$$X_{i,t} = \eta_i + \delta_i I_{t>t_0} + \epsilon_{i,t}, \quad i = 1,\ldots,N,\ t = 1,\ldots,T,$$

implying that the outcome only depends on an individual location η_i, an individual level-shift δ_i at a common time point t_0 and a random error $\epsilon_{i,t}$. In this setting we test the null hypothesis of a stationary panel

$$H_0 :\ \delta_1,\ldots,\delta_N = 0$$

against the alternative of a structural break at an unknown time point t_0 :

$$H_1 :\ \exists i \in \{1,\ldots,N\} \text{ such that } \delta_i \neq 0.$$

The individual error processes $(\epsilon_{i,t})_{t=1,\ldots,T},\ i = 1,\ldots,N$ are supposed to be stationary and independent of each other. They are allowed to have different distributions and to be short range dependent. For technical assumptions on $(\epsilon_{i,t})_{i\in\{1,\ldots,N\},\ t\in\{1,\ldots,T\}}$ see Section 4.3. To avoid identification problems we set $\text{median}(\epsilon_{i,1}) = 0,\ i = 1,\ldots,N$.

In the following paragraph we develop the proposed test procedure. If one is only interested in detecting a change in one individual i it is quite common to look at its CUSUM statistic

$$Z_T^{(i)}(x) = \frac{1}{\sqrt{T}\nu_i}\left(\sum_{t=1}^{\lfloor Tx\rfloor} X_{i,t} - \frac{\lfloor Tx\rfloor}{T}\sum_{t=1}^{T} X_{i,t}\right), \quad x \in [0,1], \tag{4.1}$$

which basically compares the mean value of the first part with that of the second for every split point. A large absolute difference for any split point indicates a structural change. If there is serial dependence the CUSUM statistic (4.1) depends on the so called long run variance

$$\nu_i^2 = \text{Var}(X_{i,1}) + 2\sum_{h=1}^{\infty}\text{Cov}(X_{i,1}, X_{i,1+h}).$$

Since (4.1) is a linear statistic it is sensitive regarding unusually small or large values. To bound the influence of outliers one can transform the observations with a so called $\Psi-$function

$$Y_{i,t} = \Psi_i\left(\frac{X_{i,t}-\mu_i}{\sigma_i}\right), \quad i = 1,\ldots,N,\ t = 1,\ldots,T$$

with $\Psi_i : \mathbb{R} \to \mathbb{R},\ \mu_i \in \mathbb{R},\ \sigma_i > 0,\ i = 1,\ldots,N$. Two examples and their impact on a short time series are shown in Figure 4.1. In change-point analysis monotone $\Psi-$functions are preferred (Hušková and Marušiaková, 2012) since redescending ones not only limit the influence of unusual values, which can be a result of a large level shift, but also can shrink them to 0 as we can can see in Figure 4.1 on the right.

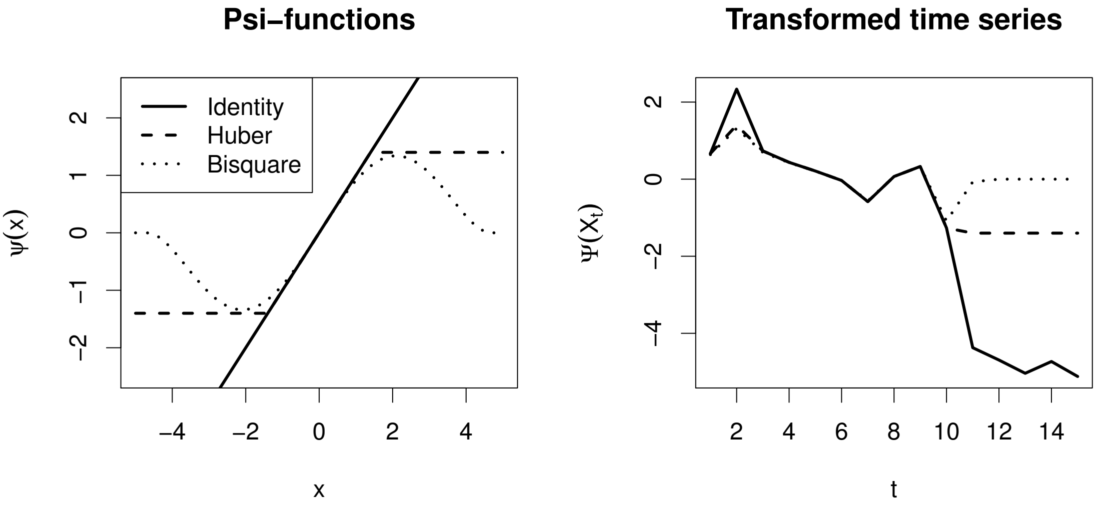

Figure 4.1: Ψ-functions (left) and corresponding transformed time series (right) with a level shift at $t = 11$.

Usually the parameters $\mu_i,\ i = 1, \ldots, N$ describe the central location and $\sigma_i,\ i = 1, \ldots, N$ the scale. Their aim is to standardize the data such that outliers are treated independently of the scale and location of the underlying distribution. Regarding the change-point problem they can also be seen as tuning parameters, since the resulting test is valid under some restrictions we will state in Section 4.3 irrespective of their particular choice. But of course some choices are more suitable than others. If σ_i is too large, outliers are hardly downweighted, and if it is too small, a lot of information is lost, see Figure 4.2. An inappropriate value of μ_i can heavily skew formerly symmetric data and even destroy the data completely, if μ_i is far away from the observed data, see Figure 4.2 on the right. We will show in Section 4.3 that under some regularity conditions we can choose μ_i and $\sigma_i,\ i = 1, \ldots, N$ data adaptively. We recommend to use highly robust and computationally fast estimators for location respectively scale, like the median and the median absolute deviation (MAD).

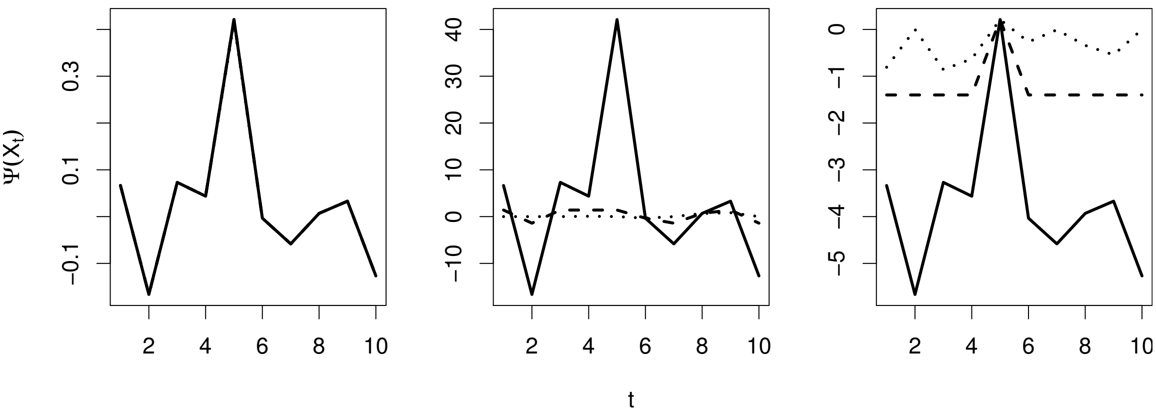

Figure 4.2: Transformed time series using the identity (solid), Huber (dashed) and bisquare (dotted) $\Psi-$function with $\mu = 0,\ \sigma = 10$ (left), $\mu = 0,\ \sigma = 0.1$ (middle) and $\mu = 4,\ \sigma = 1$ (right).

The CUSUM statistic of the transformed panel is then defined as

$$S_T^{(i)}(x) = \frac{1}{\sqrt{T}v_i}\left(\sum_{t=1}^{\lfloor Tx\rfloor} Y_{i,t} - \frac{\lfloor Tx\rfloor}{T}\sum_{t=1}^{T} Y_{i,t}\right), \quad x \in [0,1]$$

with

$$v_i^2 = \operatorname{Var}(Y_{i,1}) + 2\sum_{h=1}^{\infty}\operatorname{Cov}(Y_{i,1}, Y_{i,1+h}).$$

In the one dimensional case $N = 1$, setting $\sigma_1 = 1$ and choosing μ_1 as the M-estimator corresponding to the chosen $\Psi-$function this statistic was proposed by Hušková and Marušiaková (2012) for change point detection.
Assuming independence between the individuals the related multivariate Wald-type test for finite N equals

$$\sum_{i=1}^{N}\left(S_T^{(i)}(x)\right)^2, \quad x \in [0,1], \tag{4.2}$$

which converges for $T \to \infty$ under some regularity conditions to the sum of independent squared Brownian bridges. If additionally $N \to \infty$ one has to normalize (4.2):

$$W_{N,T}(x) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\left(S_T^{(i)}(x)\right)^2 - \frac{\lfloor xT\rfloor(T-\lfloor xT\rfloor)}{T^2}\right), \quad x \in [0,1].$$

Note that $\frac{\lfloor xT\rfloor(T-\lfloor xT\rfloor)}{T^2}$ nearly equals the variance of a Brownian bridge and therefore approximates the mean of $S_T^{(i)}(x)$ for large T. Setting $\Psi_i(x) = x,\ i = 1, \ldots, N,\ \mu_i$ cancels out, σ_i is absorbed into v_i and one arrives at the non-robust panel-CUSUM-statistic which was originally proposed by Bai (2010) and investigated theoretically by Horváth and Hušková (2012). If T tends faster to infinity than N (the accurate rates can be found in Section 4.3) and under some regularity conditions $W_{N,T}(x)$ converges weakly to a Gaussian process $\Gamma(x)$ defined by

$$\mathbb{E}(\Gamma(x)) = 0 \ \text{ and } \ \operatorname{Cov}(\Gamma(x), \Gamma(y)) = x^2(1-y^2),\ 0 \leq x \leq y \leq 1.$$

It is shown in Horváth and Hušková (2012) that such a process can be simulated based on a standard Brownian motion $(B(t))_{0\leq t<\infty}$ using the following relationship:

$$\{\Gamma(x),\ 0 \leq x \leq 1\} \overset{D}{=} \left\{\sqrt{2}(1-x)^2 B\left(\frac{x^2}{1-x^2}\right),\ 0 \leq x \leq 1\right\}.$$

One rejects the null hypothesis of a stationary panel if

$$\sup_{0<x<1}|W_{N,T}(x)| \tag{4.3}$$

exceeds a certain quantile of $\sup_{0<x<1}|\Gamma(x)|$. A small selection of critical values can be found in Table 4.1.

α	0.9	0.95	0.975	0.99	0.995
q_α	0.899	0.990	1.072	1.173	1.245

Table 4.1: Quantiles of $\sup_{0<x<1}|\Gamma(x)|$.

The proposed test statistic (4.3) has the largest power if the change occurs in the middle of the sample. If one needs higher power near the margins, one can look at other functionals of $W_{N,T}(x)$, for example $\int_0^1 |W_{N,T}(x)|\, dx$.
In practice $v_i,\ i = 1, \ldots, N$, is unknown and has to be estimated. We propose to use a kernel estimator, which originally goes back to Parzen (1957). Denote by

$$\hat{\gamma}_i(h) = \frac{1}{T} \sum_{t=1}^{N-h} \left(Y_{i,t} - \overline{Y}_i\right) \left(Y_{i,t+h} - \overline{Y}_i\right)$$

with $\overline{Y}_i = \frac{1}{T} \sum_{t=1}^{T} Y_{i,t}$ the empirical autocovariance of $(Y_{i,t})_{t=1,\ldots,T}$ of lag $h,\ h = 1, \ldots, T-1$. Let furthermore $k : \mathbb{R} \to [-1, 1]$ be a kernel function and b_T a bandwidth, then

$$\hat{v}_i^2 = \hat{\gamma}_{i,0} + 2 \sum_{h=0}^{b_T} \hat{\gamma}_i(h) k\left(\frac{h}{b_T}\right) \tag{4.4}$$

represents the related kernel estimator for v_i^2. More details and theoretical conditions on k and b_T can be found in the next section. Simulations in Section 4.4 indicate that the flat top kernel $k = k_F$

$$k_F(x) = \begin{cases} 1 & |x| \leq 0.5 \\ 2 - 2|x| & 0.5 < |x| \leq 1 \\ 0 & |x| > 1 \end{cases} \tag{4.5}$$

with $b_T = T^{0.4}$ works well if the serial dependence is not very large.

4.3 Theoretical results

In this section we give theoretical justification of the testing procedure and compile all conditions which are necessary for the asymptotical results. We start by defining the type of short range dependence we impose on the error processes $(\epsilon_{i,t})_{t=1,\ldots,T},\ i = 1, \ldots, N$. We assume that they are near epoch dependent in probability (P-NED) on an absolutely regular process.

Definition 1. *i) Let $A, B \subset F$ be two $\sigma-$fields on the probability space (Ω, F, P). Then the regularity coefficient of A and B is defined as*

$$\beta(A, B) = \mathbb{E}\left(\sup_{M \in A} |P(M|B) - P(M)|\right).$$

ii) For a stationary process $(Z_t)_{t\in\mathbb{Z}}$, the absolute regularity coefficients are given by

$$\beta_k = \sup_{n\in\mathbb{Z}} \beta(F_{n+k}^{\infty}, F_{-\infty}^{n}),$$

where $F_n^k = \sigma(Z_n, \ldots, Z_k)$ denotes the $\sigma-$field generated by $Z_n, \ldots, Z_k$. A process is called absolutely regular, if $\beta_k \to 0$ for $k \to \infty$.

iii) The process $(X_t)_{t\in\mathbb{N}}$ is P-NED on the absolutely regular process $(Z_t)_{t\in\mathbb{Z}}$, if there is a sequence of approximating constants $(a_k)_{k\in\mathbb{N}}$ fulfilling $a_k \to 0$ for $k \to \infty$, a sequence of functions $f_k : \mathbb{R}^k \to \mathbb{R}$ for $k \in \mathbb{N}$ and a decreasing function $\Phi :\ (0, \infty) \to (0, \infty)$ such that

$$P(|X_0 - f_k(Z_{-k}, \ldots, Z_k)| > \epsilon) \leq a_k \Phi(\epsilon) \tag{4.6}$$

Remark 3.

- *Looking at functionals of processes is quite natural, in fact ARMA and GARCH models are defined as functionals of iid sequences. Part iii) demands that the influence of observations of the underlying process $(Z_t)_{t\in\mathbb{Z}}$, corresponding to time points far away from the observed X_t, is small and even vanishes, if the time difference between them tends to infinity.*

- *The underlying process $(Z_t)_{t\in\mathbb{Z}}$ is sometimes assumed to be iid, see for example Hörmann (2008) and Aue et al. (2009), and though this class would yet be quite general, this assumption is in our case more restrictive than necessary. A main component of our proofs are moment inequalities, which work fine for absolutely regular processes (Philipp, 1986).*

- *The concept of near epoch dependence has a long history. It was introduced in Ibragimov (1962) where the difference between the approximating functional $f_k(Z_{-k},\dots,Z_k)$ and X_0 was measured by the L_2 norm, which requires existing second moments of $(X_t)_{t\in\mathbb{N}}$. With applications under heavy tailed distributions and robustness in mind this is somehow restrictive, which was the reason Bierens (1981) proposed to measure the approximation error in probability. This concept is often called L_0 near epoch dependence (L_0-NED) since it is the natural limit of L_p-NED for $p\to 0$, see Prucha and Pötscher (1997) p.49. We use here the variation presented by Dehling et al. (2017) which adds the error function Φ.*

- *Let Ψ be Lipschitz continuous with Lipschitz constant L, then it is easy to see that with $(X_t)_{t\in\mathbb{N}}$ also $(\Psi(X_t))_{t\in\mathbb{N}}$ is P-NED on the same underlying process $(Z_t)_{t\in\mathbb{Z}}$ with approximating functionals $\tilde{f}_k = \Psi\circ f_k$ and constants $(a_k)_{k\in\mathbb{Z}}$. For the error function corresponding to $(\Psi(X_t))_{t\in\mathbb{N}}$ we have $\tilde{\Phi}(x) = \Phi(x/L)$. Furthermore, if Ψ is bounded by a constant c, one can find approximating functions $\tilde{f}_k$ such that $\tilde{\Phi}(|2c+\epsilon|) = 0,\ \epsilon > 0$.*

We allow for different distributions and dependence structures of the individual error processes $(\epsilon_{i,t})_{t=1,\dots,T}$ for $i=1,\dots,N$. However we need certain bounds to rule out that moments diverge to infinity or dependence gets uncontrollable strong as N tends to infinity. We therefore introduce the abbreviations $\inf^\star_i a_i = \lim_{N\to\infty}\min_{i\le N} a_i$ and $\sup^\star_i a_i = \lim_{N\to\infty}\max_{i\le N} a_i$ for a sequence $(a_i)_{i\in\mathbb{N}}$. In the following we compile some assumptions on the errors $(\epsilon_{i,t})_{i=1,\dots,N,\ t=1,\dots,T}$, which we need repeatedly.

Assumption 1. *Let $(\epsilon_{i,t})_{t=1,\dots,T}$ be a stationary process, which is P-NED on absolutely regular process $(Z_{i,t})_{t\in\mathbb{Z}}$ with approximating constants $(a_{k,i})_{k\in\mathbb{N}}$, regularity coefficients $(\beta_{k,i})_{k\in\mathbb{N}}$ and error functions Φ_i for $i=1,\dots,N$, which fulfil*

I) $(\epsilon_{1,t})_{t=1,\dots,T},\dots,(\epsilon_{N,t})_{t=1,\dots,T}$ *are independent processes,*

II) ***either*** *a):*

i) $\sup^\star_i |Y_{i,1}| = c_1 < \infty$ *a.s.,*

ii) there exist $c_2 > 0$ *and* $b > 8$ *such that* $\sup^\star_i a_{k,i}, \sup^\star_i \beta_{k,i} \le c_2(1+k)^{-b}$,

iii) there exists $\Phi:\mathbb{R}_+\to\mathbb{R}_+$ *with* $\sup^\star_i \Phi_i(x) \le \Phi(x)\ \forall x > 0$ *such that* $\int_0^1 \Phi(x)\,dx < \infty$,

or *b):*

i) there exist $a \ge 8$ *such that* $\sup^\star_i \mathbb{E}\left(|Y_{i,1}|^a\right) = c_1$,

ii) there exist $c_2 > 0,\ b > 8$ *such that for* $sup^\star_i a_{k,i} \le c_2(1+k)^{-\frac{b\cdot a}{a-7}}$ *and* $\sup^\star_i \beta_{k,i} \le c_2(1+k)^{-\frac{b\cdot a}{a-8}}$,

iii) there exists $\Phi:\mathbb{R}_+\to\mathbb{R}_+$ *with* $\sup^\star_i \Phi_i(x) \le \Phi(x)\ \forall x > 0$ *such that* $\int_1^\infty x^{a-1}\Phi(x)\,dx < \infty$ *and* $\int_0^1 x^{\frac{a}{a-7}-1}\Phi(x)\,dx < \infty$.

Remark 4.

- *Assumption I) is very strong but essential in the proof. Deriving asymptotics under dependence between panels is for this test statistic if at all only possible under very restrictive assumptions. For example Horváth and Hušková (2012) add dependence through common time effects, which need to shrink with increasing N. Another possibility could be to impose a spatial dependence structure like in Jirak et al. (2015).*
- *Assumption II) differentiates depending on whether the transformed random variables $Y_{i,t}$ are bounded or not. The former is the standard case for a robust procedure, since it requires bounded $\Psi-$ functions. The later contains the non-robust test with $\Psi_i(x) = x, i = 1, \ldots, N$.*
- *If one uses finitely many bounded Ψ-functions assumption II) a) i) is always fulfilled regardless of the underlying distribution. This is a major advantage of the robust method over the non-robust one which depends on finite moments of order 8, see also Horváth and Hušková (2012).*
- *If $|\Psi_i(\epsilon_{i,1})|$ is not bounded one needs at least finite eighth moments of the transformed time series. If higher moments exist, one can weaken the dependence condition II) b) ii). Assumption II) b) iii) is quite technical and implies some finite moments of the error of the P-NED approximation. Using the P-NED concept is quite artificial in this case. Nevertheless we will prove the results also under these assumptions since it generalizes the results of Horváth and Hušková (2012) to non-linear processes.*
- *Assumption II a)/b) ii) define that the dependence needs to decay uniformly. A similar condition can also be found in Horváth and Hušková (2012).*

In the following theorem we derive the asymptotic behaviour of $W_{N,T}(x)$ if v_i is known and σ_i, μ_i are fixed for $i = 1, \ldots, N$. Let for this purpose $\overset{D[0,1]}{\to}$ denote weak convergence in the Skorokhod space $D[0, 1]$.

Theorem 3. *Let $(\epsilon_{i,t})_{i=1,\ldots,N,\ t=1,\ldots,T}$ fulfil Assumption 1 and furthermore*

i) $N, T \to \infty$ *with* $N/T \to 0$

ii) the $\Psi-$*functions are Lipschitz continuous with constants* L_i *and* $\sup_i^\star L_i = c_3 < \infty$,

iii) $\inf_i^\star v_i = \delta > 0$,

iv) $\inf_i^\star \sigma_i = \sigma_0 > 0$,

then

$$(W_{N,T(x)})_{x\in[0,1]} \overset{D[0,1]}{\to} (\Gamma(x))_{x\in[0,1]}.$$

Remark 5.

- *The special case $\Psi_i(x) = x$ is treated by Horváth and Hušková (2012). Note that the functions $(\Psi_i)_{i\in\mathbb{N}}$ do not interfere with the limiting distribution as long as they do not cut existing moments or lead to imploding variances.*
- *Condition i) prohibits N to grow faster than T. Intuitively one might think that a large N always improves the asymptotics, but one needs to remember that the standardization by $\frac{\lfloor xT\rfloor(T-\lfloor xT\rfloor)}{T^2}$ and v_i is only an approximation for large T. If N grows too fast these small errors sum up to something which is not negligible. In Horváth and Hušková (2012) only $N/T^2 \to 0$ is required and it is not clear why we need this somewhat stronger assumption here, which is by the way only necessary to verify tightness. A possible reason is our more flexible dependence structure of the errors. However, since the following Theorems also require $N/T \to 0$, we do not judge this more restrictive condition as a real drawback.*

- *Assumption ii) implies some regularity conditions for the Ψ-functions. Similar conditions can be found in one dimensional change point detection, see Hušková and Marušiaková (2012). For usual ψ functions this is no restriction.*

- *Condition iii) states that the variability within one individual should not tend to 0 and this condition can also be found in Horváth and Hušková (2012). Though one has to mention that the condition is here formulated based on the transformed time series. Inappropriate choices of μ_i and σ_i can shrink all values to a constant, see Figure 4.2, and so cause the assumption to be violated, though the original time series was permissible.*

To actually apply the testing procedure, one has to find estimators for $v_i,\ i = 1, \dots, N$. As usual the asymptotics depend on the smoothness of the kernel k in 0. We need the following assumptions on the kernel.

Assumption 2. *Let $k : \mathbb{R} \to [-1, 1]$ be a kernel function with*

i) $k(0) = 1$,

ii) $k(-x) = k(x), x \geq 0$,

iii) $k(x) = 0,\ |x| \geq 1$

iv) the first $m - 1 > 0$ derivatives of k in 0 are 0,

v) $\exists \epsilon, M > 0$ such that $|k^{(m)}(x)| \leq M$ for $|x| \leq \epsilon$, where $k^{(m)}$ denotes the $m-th$ derivative of k.

Remark 6.

- *While conditions i) and ii) are very conventional, see for example section 9.3.2. in Anderson (1971), requirement iii) is more particular, though fulfilled for most kernels.*

- *The number m characterizes the smoothness of the kernel in 0 which determines the bias of the kernel estimator, see also chapter 9.3.2 of Anderson (1971). The popular Daniell, Blackman-Tukey, Hanning, Hamming and Parzen kernels fulfil condition iv) with $m = 2$, the flat top kernel (4.5) even for arbitrary $m \in \mathbb{N}$. Since the Bartlett-Kernel is not differentiable in 0, it is not covered by the assumptions.*

- *Condition v) is fulfilled for all above mentioned kernels except the Bartlett kernel. Both assumption iv) and v) are additional requirements, not necessary in the usual one- or multidimensional time series context. We need them here to ensure uniform convergence over all individuals.*

The next theorem deals with the case where the theoretical long run variances v_i are replaced by their kernel estimations (4.4).

Theorem 4. *Let $Y_{i,t} = \Psi_i\left(\frac{X_{i,t}-\mu_i}{\sigma_i}\right)$ and denote*

$$\tilde{S}_T^{(i)}(x) = \frac{1}{\sqrt{T}\hat{v}_i}\left(\sum_{t=1}^{\lfloor Tx \rfloor} Y_{i,t} - \frac{\lfloor Tx \rfloor}{T}\sum_{t=1}^{T} Y_{i,t}\right), \quad x \in [0, 1],$$

the CUSUM-statistic with estimated long run variance using (4.4) and

$$\tilde{W}_{N,T}(x) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\left(\tilde{S}_T^{(i)}(x)\right)^2 - \frac{\lfloor xT \rfloor (T - \lfloor xT \rfloor)}{T^2}\right), \quad x \in [0, 1],$$

the related panel-CUSUM-statistic. Let additionally to the Assumptions of Theorem 3

$$Nb_T/T \to 0 \text{ and } N/b_T^{2s} \to 0 \quad \text{where} \quad s = \begin{cases} \min(m, b-1) & m \neq b-1 \\ m-1 & m = b-1 \end{cases}, \tag{4.7}$$

then

$$\sup_{x\in[0,1]} |\tilde{W}_{N,T}(x) - W_{N,T}(x)| \xrightarrow{p} 0.$$

Remark 7.

- *The assumption $Nb_T/T \to 0$ seems to be stronger than necessary. In the same situation Horváth and Hušková (2012) only require $Nb_T^2/T^2 \to 0$. The reason seems to be a difference in the proofs. While at one point Horváth and Hušková (2012) consider the L_2 norm, we choose the L_1 norm to make the proof somewhat more feasible. Remember that we allow for a little more complicated dependence structure.*
- *As usual for kernel estimators the rate of b_T should neither be too fast (large variance) nor too slow (large bias). This is reflected in (4.7). Furthermore, we see that from the theoretical point of view the flat-top kernel is preferable. In this case the bias condition $N/b_T^{2s} \to 0$ only depends on the strength of the serial dependence.*

Now we want to prove that the tuning parameters μ_i and σ_i, $i = 1\ldots, N$ can be chosen data-adaptively. Let therefore $\mu(\cdot)$ denote a univariate location measure, which implies $\mu(F^\star) = a + \mu(F)$ for $a, b \in \mathbb{R}$ and any one-dimensional distribution F where $F^\star$ is the distribution of $a + bX$ with $X \sim F$. Examples are the mean $\mu_{mean}(F) = \mathbb{E}(X)$ and the more robust median $\mu_{med}(F) = \text{median}(X)$. For a univariate scale measure $\sigma(\cdot)$ we demand $\sigma(F^\star) = |b|\sigma(F)$. Maybe the most popular representative is the standard deviation $\sigma_{SD}(F) = \sqrt{\mathbb{E}(X - \mathbb{E}(X))^2}$, which is of course not robust. A more appropriate choice here is the median absolute deviation $\sigma_{MAD}(F) = c_F \cdot \text{median}(|X - \text{median}(X)|)$, where often $c_F = 1.4826$ so that $\sigma_{SD}(F) = \sigma_{MAD}(F)$ if F is a normal distribution.

The related estimators $\hat{\mu}$ and $\hat{\sigma}$ are usually the measure $\mu(\cdot)$ respectively $\sigma(\cdot)$ applied to the empirical distribution $\hat{F}_T$ of the sample $(X_1, \ldots, X_T)$. Though there are sometimes differences. The standard deviation is for example often defined as $\hat{\sigma}_{SD}(\hat{F}_T) = \sqrt{\frac{1}{T-1}\sum_{i=1}^T (X_i - \frac{1}{T}\sum_{i=1}^T X_i)^2}$ instead of $\hat{\sigma}_{SD}(\hat{F}_T) = \sqrt{\frac{1}{T}\sum_{i=1}^T (X_i - \frac{1}{T}\sum_{i=1}^T X_i)^2}$.

From now on we define the standardization parameters μ_i and σ_i as $\mu_i = \mu(F_i)$ and $\sigma_i = \sigma(F_i)$, $i = 1, \ldots, N$, where F_i is the marginal distribution of $X_{1,t}$ for some location measure $\mu(\cdot)$ and scale measure $\sigma(\cdot)$. The corresponding estimators are denoted by $\hat{\mu}_{i,T}$ and $\hat{\sigma}_{i,T}$ for $i = 1, \ldots, N$. The next Theorem gives conditions under which it is asymptotically negligible whether one knows these theoretical values or estimates them, as long as the estimators converge fast enough.

Theorem 5. *Denote by*

$$\check{S}_T^{(i)}(x) = \frac{1}{\sqrt{T}\check{v}_i}\left(\sum_{t=1}^{\lfloor Tx\rfloor} \Psi_i\left(\frac{X_{i,t} - \hat{\mu}_{i,T}}{\hat{\sigma}_{i,T}}\right) - \frac{\lfloor Tx\rfloor}{T}\sum_{t=1}^{T} \Psi_i\left(\frac{X_{i,t} - \hat{\mu}_{i,T}}{\hat{\sigma}_{i,T}}\right)\right), \quad x \in [0,1],$$

the CUSUM-statistic with estimated location and scale parameter, $\check{v}_i$ the long run variance estimation based on $\hat{\mu}_{i,T}$ and $\hat{\sigma}_{i,T}$ and

$$\check{W}_{N,T}(x) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\left(\check{S}_T^{(i)}(x)\right)^2 - \frac{\lfloor xT\rfloor(T - \lfloor xT\rfloor)}{T^2}\right), \quad x \in [0,1].$$

the related panel-CUSUM-statistic. Let additionally to the Assumptions of Theorem 4 hold that *i)* Ψ_i *is* $m+1$ *times continuously differentiable with derivatives* $\Psi_i^{(1)}, \dots, \Psi_i^{(m+1)}$ *and* $\sup_i^\star |\Psi_i^{(k)}(x)x^k|_\infty = c_k < \infty$, *for* $k = 1, \dots, m$,

ii) there exist $\epsilon, \delta > 0$ *such that* $\sup_i^\star \sup_{x \in \mathbb{R}, |y| \leq \delta, |z| \leq \epsilon} |\Psi_i^{(m+1)}(x(1+y)+z)x^k| = d_k < \infty$, *for* $k \leq m+1$,

iii) there exists $\alpha, \beta > 0$ *such that for* $\forall x > 0$*:*

$$\limsup_{T\to\infty} T^\alpha \overset{\star}{\sup_i} P\left(|\hat{\mu}_{i,T} - \mu_i| > \frac{x}{T^\beta}\right) < \infty \text{ and } \limsup_{T\to\infty} T^\alpha \overset{\star}{\sup_i} P\left(|\hat{\sigma}_{i,T} - \sigma_i| > \frac{x}{T^\beta}\right) \leq \infty,$$

iv) $N/T^{2\beta} \to 0$, $N/T^\alpha \to 0$ *and* $N/T^{2\beta(m+1)-1} \to 0$,

then

$$\sup_{x\in[0,1]} |\check{W}_{N,T}(x) - W_{N,T}(x)| \xrightarrow{p} 0.$$

Remark 8.

- *Assumption i) is often not fulfilled for standard* $\Psi-$*functions. The Huber* $\Psi-$*function is for example only continuous but in two points not differentiable. This assumption may be relaxed to the case where the function is differentiable in all but a finite number of points at least as the error distribution is continuous. This would be fulfilled for all common* $\Psi-$*functions. Nevertheless the current condition is not a restriction in practice, since one can always find* $m+1$*-times differentiable modifications which hardly differ from the original* $\Psi-$*function. The existence of the upper bounds* c_k *for* $k = 1, \dots, m$ *and* d_k *for* $k = 0, \dots, m+1$ *are also no substantial restrictions, since* $\Psi-$*functions are usually constant for large values (and so its derivatives are 0).*
- *Assumption iii) is the main restriction of the Theorem. We will show in the next Theorem that one can find tail probability bounds for median and MAD which imply this condition. Note that the parameter* α *and* β *determine the rate of convergence as can be seen in Assumption iv).*
- *Additionally to the persisting conditions on the ratio of* N *and* T*, we get the assumptions* $N/T^{2\beta} \to 0$, $N/T^\alpha \to 0$ *and* $N/T^{\beta(m+1)-1} \to 0$. *If exponential inequalities for* $\hat{\mu}_{i,T}$ *and* $\hat{\sigma}_{i,T}$, $i = 1, \dots, N$ *are available and* Ψ_i *is 2 times continuously differentiable,* $i = 1, \dots, N$ *this boils down to* $N/T^{1-\epsilon} \to 0$ *for some* $\epsilon > 0$. *So* T *has to grow a little faster than* N.

Finally we want to investigate if our proposed estimators median and MAD fulfil assumption *iii)* of Theorem 5 and how the parameters α and β depend on the properties of the observed processes. We formulate the result for a one dimensional time series $(X_t)_{t\in\mathbb{N}}$ respectively its transformation $(Y_t)_{t\in\mathbb{N}}$ with $Y_t = |X_t - \mu_{med}|$, $t \in \mathbb{N}$.

Theorem 6. *Let* $(X_t)_{t\in\mathbb{N}}$ *be stationary and P-NED on an absolutely regular process* $(Z_t)_{t\in\mathbb{Z}}$ *with approximation constants* $(a_k)_{k\in\mathbb{N}}$, *functions* $(f_k)_{k\in\mathbb{N}}$, *error function* Φ *and absolutely regularity coefficients* $(\beta_k)_{k\in\mathbb{N}}$. *Furthermore*

i) there exist $\kappa > 0$ *and* $p \in \mathbb{N}$ *even such that* $\sum_{k=1}^\infty (a_k\Phi(a_k^\kappa) + a_k^\kappa)k^p < \infty$ *and* $\sum_{k=1}^\infty \beta_k k^p < \infty$,

ii) X_1 *is continuous with bounded density* f *and there exist* $M, \epsilon > 0$ *with* $f(x) \geq M$ *for* $x \in (\mu_{med} - \epsilon, \mu_{med} + \epsilon)$,

then there exists c *such that for* $x > 0$

$$P(|\hat{\mu}_{med,T} - \mu_{med}| > xT^{-\beta}) \leq \frac{c}{x^p} T^{-p/2+p\beta} \; \forall T \in \mathbb{N}. \tag{4.8}$$

If additionally

iii) Y_1 is continuous with bounded density g and there exist $M, \epsilon > 0$ with $g(x) \geq M$ for $x \in (\sigma_{MAD}/c_F - \epsilon, \sigma_{MAD}/c_F + \epsilon)$,

then there exists c such that for $x > 0$

$$P(|\hat{\sigma}_{MAD,T} - \sigma_{MAD}| > xT^{-\beta}) \leq \frac{c}{x^p} T^{-p/2+p\beta} \; \forall T \in \mathbb{N}.$$

Remark 9.

- *Under independence there exist exponential inequalities for the median and the MAD, see Serfling and Mazumder (2009). We are not aware of such exponential inequalities under dependence, but looking at the proof in Serfling (1980) or Serfling and Mazumder (2009) it only depends on the existence of a Hoeffding inequality, which is proven under various dependence conditions, see Kontorovich and Ramanan (2008) or Kallabis et al. (2006). However, proving such inequalities under the P-NED condition is not the objective of this thesis.*
- *The condition of a continuous distribution might be relaxed if one slightly redefines median and MAD, see also Serfling and Mazumder (2009).*

4.4 Simulation

In this section we want to evaluate our proposed test statistic concerning two aspects: The size under the null hypothesis and the power under the alternative. There are quite different simulation scenarios possible, since we allow for quite diverse serial dependence structures and distributions of the individuals. For the reason of comparison we orientate ourselves at the AR(1) models considered in Horváth and Hušková (2012)

$$X_{i,t} = \rho \cdot X_{i,t-1} + a_{i,t}, \; t = 1, \ldots, T, \; i = 1, \ldots, N.$$

We compare the non-robust panel CUSUM statistic proposed by Horváth and Hušková (2012) with our robust alternative. We use a two times continuously differentiable version of Hubers-$\Psi-$function which is shown in Figure 4.3. Furthermore we choose the flat top kernel (4.5) with a bandwidth $b_T = T^{0.4}$ for both estimators. Note that Horváth and Hušková (2012) originally use a rather small value of b_T varying between 2.5 and 5 which is, however, not appropriate in our simulations where we also look at relatively large T. Simulation results are based on 1000 runs each.

We first look at finite sample properties under the null hypothesis. In the case of $\rho = 0$, where we have no serial dependence, both tests need at least $T = 200$ to work properly as one can see in the upper quarter of Table 4.2. Furthermore the simulations confirm the theoretical results that N must not grow much faster than T. We also notice that there is basically no difference between the tests under Gaussianity. In case of a more heavy tailed distribution like a t-distribution with 3 degrees of freedom, results do not change much, as one can see in the third quarter of Table 4.2. The non-robust test seems to work well under the null hypothesis though its assumptions (finite 8-th moments) are violated. Somewhat surprisingly the results get better under serial correlation ($\rho = 0.5$) where we get reasonable empirical sizes already for $T = 100$ for normal as well as t_3 distributed innovations. This can be partly explained by the relatively large bandwidth which better fits the case $\rho = 0.5$ than $\rho = 0$.

It is maybe preferable to use a data dependent bandwidth b_T as proposed in Andrews (1991) or Politis (2011). In this case one can either calculate a global bandwidth which is used for every individual or customized bandwidths $b_{T,i}, \; i = 1, \ldots, N$. For the latter one has to ensure that the conditions (4.7) are also fulfilled for the infimum respectively the supremum of the bandwidths. However, automatic bandwidth choices are also questioned since they can produce very large bandwidths under the alternative of a level shift which considerably decrease the power of

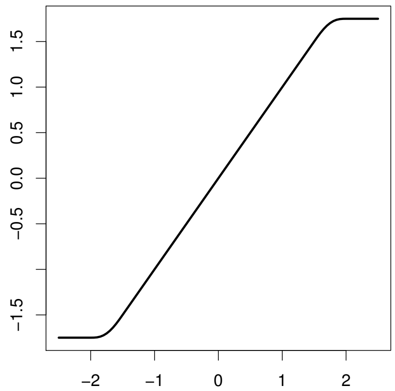

Figure 4.3: Two times continuously differentiable version of Hubers $\Psi-$function used in the simulations.

			non-robust test					robust test				
	ρ	N T	50	100	200	400	800	50	100	200	400	800
N	0	50	0.33	0.13	0.07	0.07	0.04	0.33	0.14	0.07	0.06	0.04
		100	0.52	0.14	0.06	0.07	0.06	0.55	0.15	0.07	0.06	0.06
		200	0.77	0.21	0.11	0.07	0.04	0.79	0.22	0.11	0.08	0.04
		400	0.97	0.40	0.18	0.08	0.08	0.97	0.41	0.17	0.07	0.08
		800	1.00	0.65	0.28	0.13	0.06	1.00	0.66	0.28	0.13	0.07
	0.5	50	0.07	0.06	0.05	0.04	0.03	0.08	0.05	0.05	0.05	0.03
		100	0.12	0.07	0.04	0.04	0.05	0.12	0.07	0.04	0.04	0.06
		200	0.24	0.10	0.05	0.04	0.04	0.26	0.11	0.05	0.05	0.04
		400	0.53	0.16	0.07	0.06	0.04	0.54	0.17	0.07	0.05	0.04
		800	0.90	0.36	0.13	0.07	0.05	0.90	0.37	0.13	0.07	0.05
t_3	0	50	0.30	0.11	0.05	0.08	0.06	0.34	0.10	0.05	0.08	0.06
		100	0.47	0.15	0.08	0.06	0.06	0.54	0.16	0.08	0.06	0.06
		200	0.74	0.19	0.10	0.06	0.05	0.77	0.25	0.11	0.07	0.06
		400	0.97	0.36	0.12	0.08	0.06	0.97	0.41	0.15	0.09	0.06
		800	1.00	0.65	0.23	0.14	0.08	1.00	0.68	0.26	0.15	0.08
	0.5	50	0.07	0.04	0.05	0.04	0.05	0.08	0.04	0.04	0.04	0.05
		100	0.11	0.05	0.03	0.04	0.04	0.10	0.06	0.05	0.06	0.04
		200	0.25	0.10	0.05	0.04	0.04	0.26	0.10	0.06	0.05	0.04
		400	0.52	0.19	0.06	0.04	0.04	0.54	0.20	0.07	0.05	0.05
		800	0.90	0.38	0.12	0.07	0.05	0.89	0.39	0.14	0.07	0.06

Table 4.2: Empirical size under normal (N) and t_3 distributed innovations, different AR(1) parameters ρ, time dimensions T, number of individuals N and a significance level of 0.05.

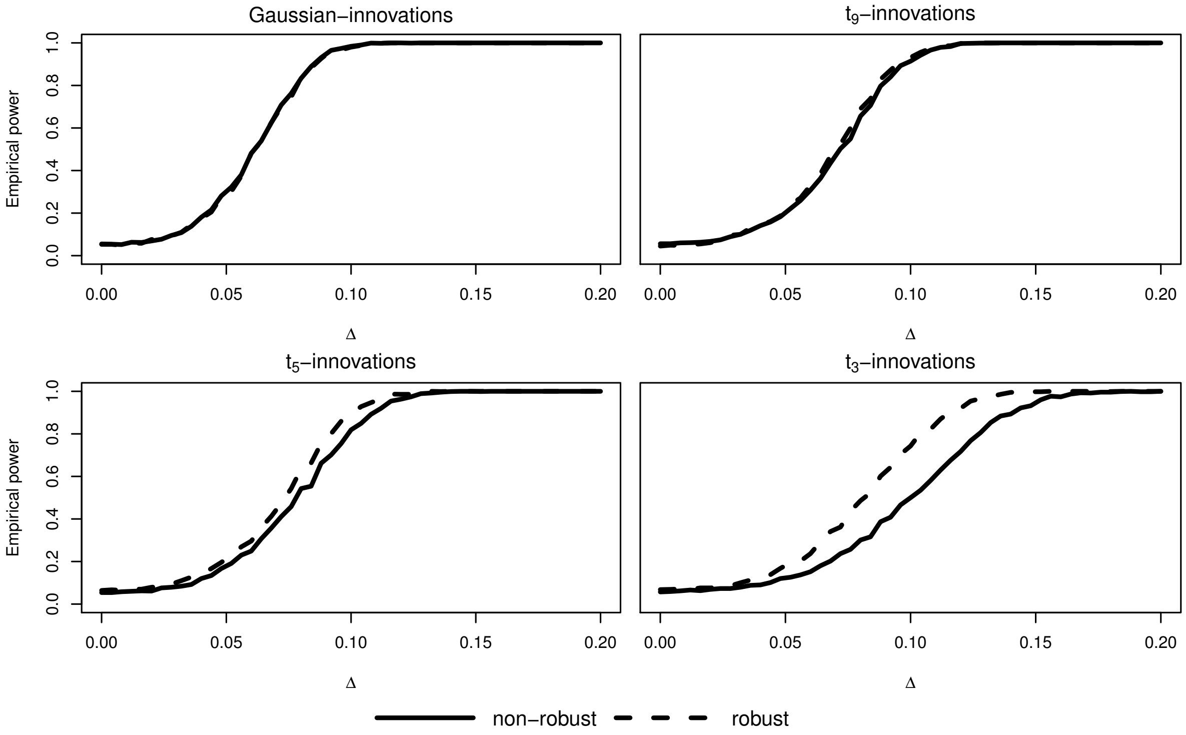

Figure 4.4: Empirical power under $\rho = 0.25$, different distributed innovations, $N = 200$, $T = 400$ and a jump at $T = 200$ which is generated by a normal distribution with mean 0 and standard deviation Δ.

CUSUM tests. They can even lead to non-monotonic power curves, see for example Vogelsang (1999) and Crainiceanu and Vogelsang (2007).
Now we turn our focus at the behaviour under the alternative. Therefore we choose $\rho = 0.25$, set $T = 400$, $N = 200$ and add a jump at $t_0 = 200$. Its height is generated for each individual independently by a normal distribution with mean 0 and standard deviation Δ varying between 0 and 0.2. Results for normal and t-distributed innovations with various degrees of freedom can be found in Figure 4.4. We see that there is no visual difference between the tests under the alternative in case of normal data. This changes as the innovations get more heavy tailed. The robust test statistic performs superior in case of t_5 distributed innovations. This advantage increases if one looks at more heavy tailed distributions. Under t_1 innovations the power of the non-robust test even does not exceed its size. To sum up, the robust test performs comparably to the non-robust one in the Gaussian case and clearly outperforms it under heavy tailed distributions.

4.5 Summary

We have proposed a robust test for change-point detection in panel data where the number of individuals and the time horizon is large. The procedure is based on residuals which are robustly transformed via a Ψ-function. The null distribution is derived under very weak conditions, allowing for arbitrarily heavy tailed distributions and heterogeneous serial dependence. To the best of our knowledge this is the first contribution in the robust change-point literature considering a data dependent choice of the tuning parameters $\sigma_i, i = 1, \ldots, N$, which allows to combine high robustness and efficiency.

A simulation study illustrates that the robust procedure outperforms the non-robust one under heavy tails and moreover indicates that there is little loss under Gaussian data. This at first glance surprising observation is not unusual for robust methods applied in the high dimensional setting. For instance, it is observed that the Pitman asymptotic efficiency under Gaussianity of the spatial median approaches 1 if the dimension N is large (Möttönen et al., 1997). Asymptotic variances of robust scatter estimators like the Tyler shape matrix, MCD, M- and S-estimators tend to that of the empirical covariance matrix under multivariate normal distributions for large N (see e.g. Croux and Haesbroeck, 1999; Taskinen et al., 2006). Relative efficiencies of sign tests for uniformity on the unit sphere, independence against serial dependence and multivariate independence tend to 1 under normality and growing dimension compared to classical Gaussian competitors (see e.g. Paindaveine et al., 2016). With regard to these results it is interesting to know whether one can prove similar results in the panel context.
There are also some limitations of our testing procedure. First there are no covariates (apart from an individual mean) in the considered model. The test statistic uses robustly transformed observations based on prior location- and scale-estimates. The intuitive generalisation would be to substitute regression residuals (from a robust regression) for the standardized residuals. However it is not clear under which conditions the asymptotic distribution is still the same as in the much simpler panel model considered here. Another limitation is the exclusion of cross sectional dependence. Promising in this regard looks the projection approach proposed by Aston and Kirch (2014), which allows for an arbitrary dependence structure between the individuals. The main challenge here, also in the non-robust case, is the choice of the projection, since the power of the test crucially depends on it. Finally our test requires that the time dimension T is large. It enables us to allow under some regularity conditions for arbitrary serial dependence which can even differ between the individuals. If T is small one has to be more restrictive in this regard and it is also more complicated to allow for different distributions of the particular individuals.

4.6 Proofs

The main component of our proofs are moment inequalities, which are based on coupling arguments. More in detail we use the following result by Philipp (1986) which describes how dependent random variables can be substituted by independent ones.

Proposition 6. *(Theorem 3.4 in Philipp (1986)): Let $\{B_k, m_k, k \geq 1\}$ be a sequence of Polish spaces. Let α_k denote the Borel field over B_k, let $\{X_k,\ k \geq 0\}$ be a sequence of random variables with values in B_k and let $\{\gamma_k,\ k \geq 0\}$ be a sequence of non-decreasing $\sigma-$fields such that X_k is γ_k-measurable. Suppose that for some sequence $\{\beta_k,\ k \geq 0\}$ of non-negative numbers*

$$\mathbb{E} \sup_{A \in \alpha_k} \left(|P(X_k \in A|\gamma_{k-1}) - P(X_k \in A)|\right) \leq \beta_k$$

for all $k \geq 1$. Denote by F_k the distribution of X_k and let $\{G_k,\ k \geq 0\}$ be a sequence of distributions on (B_k, α_k) such that

$$F_k(A) \leq G_k(A^{\rho_k}) + \sigma_k \ \ \forall\ A \in \alpha_k$$

with $\rho_k,\ \sigma_k \geq 0$ and $A^\epsilon = \cup_{x \in A}\{y :\ \ m_k(x, y) \leq \epsilon\}$. Then without changing its distribution, one can redefine the sequence $\{X_k,\ k \geq 0\}$ on a richer probability space on which there exists a sequence $\{Y_k,\ k \geq 1\}$ of independent random variables with distribution G_k such that for all $k \geq 1$:

$$P(m_k(X_k, Y_k) > \rho_k) \leq \beta_k + \sigma_k.$$

We use Proposition 6 to derive covariance inequalities for processes which are P-NED. Similar inequalities have already been proved for processes which are strong mixing (see Davydov, 1970; Deo, 1973) or L_1-NED (see Borovkova et al., 2001).

Proposition 7. *Let $(X_t)_{t\in\mathbb{N}}$ be stationary and P-NED on an absolutely regular process $(Z_t)_{t\in\mathbb{Z}}$ with approximation constants $(a_k)_{k\in\mathbb{N}}$, functions $(f_k)_{k\in\mathbb{N}}$, error function Φ and absolutely regularity coefficients $(\beta_k)_{k\in\mathbb{N}}$. There furthermore exists $a > p \geq 2$ with $\mathbb{E}(|X_1|^a) < \infty$, $\int_0^1 x^{\frac{a}{a-p+1}-1}\Phi(x)dx < \infty$, $\int_1^\infty x^{a-1}\Phi(x)dx < \infty$, then there are D_1 and D_2 independent of m such that*

$$|\mathbb{E}(X_{i_1}\dots X_{i_k}X_{i_{k+1}}\dots X_{i_p}) - \mathbb{E}(X_{i_1}\dots X_{i_k})\mathbb{E}(X_{i_{k+1}}\dots X_{i_p})| \leq D_1 a_{\lfloor m/3\rfloor}^{\frac{a-p+1}{a}} + D_2\beta_{\lfloor m/3\rfloor}^{\frac{a-p}{a}} \tag{4.9}$$

where $1 \leq i_1 \leq \dots \leq i_p \leq T$ and $m = i_{k+1} - i_k$.

Proof of Proposition 7. The proof follows the ideas of Borovkova et al. (2001). First we build independent blocks $\tilde{W}_1 = \tilde{Z}_{i_1-\lfloor m/3\rfloor},\dots,\tilde{Z}_{i_k+\lfloor m/3\rfloor}$ and $\tilde{W}_2 = \tilde{Z}_{i_{k+1}-\lfloor m/3\rfloor},\dots,\tilde{Z}_{i_p+\lfloor m/3\rfloor}$ where the functions f_m will work. Denote the original blocks by $W_1 = Z_{i_1-\lfloor m/3\rfloor},\dots,Z_{i_k+\lfloor m/3\rfloor}$ and $W_2 = Z_{i_{k+1}-\lfloor m/3\rfloor},\dots,Z_{i_p+\lfloor m/3\rfloor}$. We apply Proposition 6 with $X_1 = W_1$ and $X_2 = W_2$ as well as $\rho_k = \sigma_k = 0$ for $k = 1,2$. Then $B_1 = \mathbb{R}^{i_k-i_1+2\lfloor m/3\rfloor+1}$ and $B_2 = \mathbb{R}^{i_p-i_{k+1}+2\lfloor m/3\rfloor+1}$ are polish spaces, $\gamma_1 = \sigma(\dots,Z_{i_k+\lfloor m/3\rfloor})$, $\gamma_2 = \sigma(\dots,Z_{i_p+\lfloor m/3\rfloor})$ and

$$\mathbb{E}\sup_{A\in\alpha_r}(|P(W_r\in A|\gamma_{r-1}) - P(W_r\in A)|) \leq \beta_{m-2\lfloor m/3\rfloor} \leq \beta_{\lfloor m/3\rfloor}\ r = 1,2$$

since both blocks are separated by $m - 2\lfloor m/3\rfloor$. Proposition 6 then guarantees the existence of independent blocks $\tilde{W}_1$ and $\tilde{W}_2$ which are distributed as W_1 and W_2 such that:

$$P(\tilde{W}_i \neq W_i) \leq \beta_{\lfloor m/3\rfloor} \text{ for } i = 1,2$$

which entails that the corresponding transformations $f_{\lfloor m/3\rfloor}(\tilde{Z}_{l-\lfloor m/3\rfloor},\dots,\tilde{Z}_{l+\lfloor m/3\rfloor}) =: \tilde{X}_l$ and $f_{\lfloor m/3\rfloor}(\tilde{Z}_{n-\lfloor m/3\rfloor},\dots,\tilde{Z}_{n+\lfloor m/3\rfloor}) =: \tilde{X}_n$ are also independent as long as $l \leq i_k$ and $n \geq i_{k+1}$. In the following we need a bound for the error between $\tilde{X}_l$ and X_l for $l \in i_1,\dots,i_p$. For $i_1 \leq l \leq i_k$ one gets the following decomposition of the error

$$\mathbb{E}(|X_l - \tilde{X}_l|^{\frac{a}{a-p+1}}) = \mathbb{E}(|(X_l - \tilde{X}_l)I_{W_1=\tilde{W}_1}|^{\frac{a}{a-p+1}}) + \mathbb{E}(|(X_l - \tilde{X}_l)I_{W_1\neq\tilde{W}_1}|^{\frac{a}{a-p+1}}). \tag{4.10}$$

For the first term one can use the P-NED property iii)

$$\mathbb{E}(|(X_l - \tilde{X}_l)I_{W_1=\tilde{W}_1}|^{\frac{a}{a-p+1}}) \leq a_{\lfloor m/3\rfloor}\int_0^\infty \epsilon^{\frac{a}{a-p+1}-1}\Phi(\epsilon)d\epsilon \leq a_{\lfloor m/3\rfloor}\cdot C_1$$

where C_1 only depends on Φ. Using the c_r inequality (see Loève, 1977, p. 157) and Hölder inequality we get

$$\begin{aligned}
&\mathbb{E}(|(X_l - \tilde{X}_l)I_{W_1\neq\tilde{W}_1}|^{\frac{a}{a-p+1}})\\
&\leq 2^{\frac{a}{a-p+1}}\left(\mathbb{E}(|X_lI_{W_1\neq\tilde{W}_1}|^{\frac{a}{a-p+1}}) + \mathbb{E}(|\tilde{X}_lI_{W_1\neq\tilde{W}_1}|^{\frac{a}{a-p+1}})\right)\\
&\leq 2^{\frac{a}{a-p+1}}\left([\mathbb{E}(|X_l|^a)]^{\frac{1}{a-p+1}}P(W_1\neq\tilde{W}_1)^{\frac{a-p}{a-p+1}} + [\mathbb{E}(|\tilde{X}_l|^a)]^{\frac{1}{a-p+1}}P(W_1\neq\tilde{W}_1)^{\frac{a-p}{a-p+1}}\right).\\
&\leq 2^{\frac{a}{a-p+1}}\beta_{\lfloor m/3\rfloor}^{\frac{a-p}{a-p+1}}\left([\mathbb{E}(|X_l|^a)]^{\frac{1}{a-p+1}} + [\mathbb{E}(|\tilde{X}_l|^a)]^{\frac{1}{a-p+1}}\right)
\end{aligned}$$

and since $\tilde{W}_1$ is distributed as W_1 we have

$$\begin{aligned}\mathbb{E}(|\tilde{X}_l|^a) &= \mathbb{E}(|f_{\lfloor m/3\rfloor}(Z_{l-\lfloor m/3\rfloor},\dots,Z_{l-\lfloor m/3\rfloor})|^a)\\ &\le 2^{a-1}[\mathbb{E}(|f(Z_{l-\lfloor m/3\rfloor},\dots,Z_{l-\lfloor m/3\rfloor})-X_l|^a)+\mathbb{E}(|X_l|^a)]\\ &\le 2^{a-1}(a_{\lfloor m/3\rfloor}C_2+\mathbb{E}(|X_l|^a)).\end{aligned} \tag{4.11}$$

Therefore (4.10) is bounded by

$$\mathbb{E}(|X_l-\tilde{X}_l|^{\frac{a}{a-p+1}}) \le C_3(a_{\lfloor m/3\rfloor}+\beta_{\lfloor m/3\rfloor}^{\frac{a-p}{a-p+1}}) \tag{4.12}$$

which also holds for $i_{k+1}\le l\le i_p$. To derive covariance inequalities, we need a bound for the error between products of random variables and its copies

$$\begin{aligned}&\mathbb{E}(|X_{i_1}\dots X_{i_p}-\tilde{X}_{i_1}\dots\tilde{X}_{i_p}|) \qquad (4.13)\\ &\le \mathbb{E}(|(X_{i_1}-\tilde{X}_{i_1})(X_{i_2}\dots X_{i_p})|)+\sum_{j=2}^{p-1}\mathbb{E}(|\tilde{X}_{i_1}\dots\tilde{X}_{i_{j-1}}(X_{i_j}-\tilde{X}_{i_j})X_{i_{j+1}}\dots X_{i_p}|)\\ &+\mathbb{E}(|(\tilde{X}_{i_1}\dots\tilde{X}_{i_{p-1}})(X_{i_p}-\tilde{X}_{i_p})|).\end{aligned}$$

Using the generalized Hölder inequality and (4.12) one can bound the first summand on the right hand-side of (4.13) by

$$\begin{aligned}&\mathbb{E}(|(X_{i_1}-\tilde{X}_{i_1})(X_{i_2}\dots X_{i_p})|)\\ &\le \left[\mathbb{E}\left(|(X_{i_1}-\tilde{X}_{i_1})|^{\frac{a}{a-p+1}}\right)\right]^{\frac{a-p+1}{a}}\left[\mathbb{E}\left(|X_{i_2}|^a\right)\right]^{\frac{1}{a}}\cdots\left[\mathbb{E}\left(|X_{i_p}|^a\right)\right]^{\frac{1}{a}}\\ &\le C_4(a_{\lfloor m/3\rfloor}+\beta_{\lfloor m/3\rfloor}^{\frac{a-p}{a-p+1}})^{\frac{a-p+1}{a}}.\end{aligned}$$

Similar bounds for the other summands in (4.13) can be derived using (4.12) and (4.11) which results eventually in

$$\mathbb{E}(|X_{i_1}\dots X_{i_p}-\tilde{X}_{i_1}\dots\tilde{X}_{i_p}|) \le C_5 a_{\lfloor m/3\rfloor}^{\frac{a-p+1}{a}}+C_6\beta_{\lfloor m/3\rfloor}^{\frac{a-p}{a}}. \tag{4.14}$$

Analogously one gets

$$\mathbb{E}(|X_{i_1}\dots X_{i_k}-\tilde{X}_{i_1}\dots\tilde{X}_{i_k}|) \le C_7 a_{\lfloor m/3\rfloor}^{\frac{a-k+1}{a}}+C_8\beta_{\lfloor m/3\rfloor}^{\frac{a-k}{a}} \tag{4.15}$$

and

$$\mathbb{E}(|X_{i_{k+1}}\dots X_{i_p}-\tilde{X}_{i_{k+1}}\dots\tilde{X}_{i_p}|) \le C_9 a_{\lfloor m/3\rfloor}^{\frac{a-p+k+1}{a}}+C_{10}\beta_{\lfloor m/3\rfloor}^{\frac{a-p+k}{a}}. \tag{4.16}$$

Finally we prove the covariance inequality (4.9). Denote $A=X_{i_1}\dots X_{i_k}$, $B=X_{i_{k+1}}\dots X_{i_p}$, $\tilde{A}=\tilde{X}_{i_1}\dots\tilde{X}_{i_k}$ and $\tilde{B}=\tilde{X}_{i_{k+1}}\dots\tilde{X}_{i_p}$ then we get by (4.14), (4.15) and (4.16)

$$\begin{aligned}|\mathbb{E}(AB)-\mathbb{E}(A)\mathbb{E}(B)| &= |\mathbb{E}(AB)-\mathbb{E}(A-\tilde{A}+\tilde{A})\mathbb{E}(B-\tilde{B}+\tilde{B})|\\ &\le \mathbb{E}(|AB-\tilde{A}\tilde{B}|)+\mathbb{E}(|\tilde{A}|)\mathbb{E}(|B-\tilde{B}|)\\ &+\mathbb{E}(|\tilde{B}|)\mathbb{E}(|A-\tilde{A}|)+\mathbb{E}(|A-\tilde{A}|)\mathbb{E}(|B-\tilde{B}|)\\ &\le C_{11}a_{\lfloor m/3\rfloor}^{\frac{a-p+1}{a}}+C_{12}\beta_{\lfloor m/3\rfloor}^{\frac{a-p}{a}},\end{aligned}$$

which completes the proof. □

The result is a little sharper if the process is bounded.

Proposition 8. *Let $(X_t)_{t\in\mathbb{N}}$ be stationary and P-NED on an absolutely regular process $(Z_t)_{t\in\mathbb{Z}}$ with approximation constants $(a_k)_{k\in\mathbb{N}}$, functions $(f_k)_{k\in\mathbb{N}}$, error function Φ and absolutely regularity coefficients $(\beta_k)_{k\in\mathbb{N}}$. There furthermore exists D such that $|X_1| \leq D$ a.s. and $\int_0^\infty \Phi(x)dx < \infty$, then there exists D_1 independent of m such that*

$$|\mathbb{E}(X_{i_1}\dots X_{i_k}X_{i_{k+1}}\dots X_{i_p}) - \mathbb{E}(X_{i_1}\dots X_{i_k})\mathbb{E}(X_{i_{k+1}}\dots X_{i_p})| \leq D_1(a_{\lfloor m/3\rfloor} + \beta_{\lfloor m/3\rfloor}),$$

where $1 \leq i_1 \leq \ldots \leq i_p \leq T$ and $m = i_{k+1} - i_k$.

The proof is analogous to that of Proposition 7. But instead of using the Hölder inequality one uses the boundedness of X_i, which enables us to extract the largest possible absolute value D out of the expectation.
The next theorem covers the case where one looks at transformations of the P-NED process. It is necessary for the consistency of the long run variance estimation under estimated tuning parameters.

Proposition 9. *Let $(X_t)_{t\in\mathbb{N}}$ be stationary and P-NED on an absolutely regular process $(Z_t)_{t\in\mathbb{Z}}$ with approximation constants $(a_k)_{k\in\mathbb{N}}$, functions $(f_k)_{k\in\mathbb{N}}$, absolutely regularity coefficients $(\beta_k)_{k\in\mathbb{N}}$ and error function Φ with $\int_0^\infty \Phi(x)dx < \infty$. Furthermore let $g_1,\ldots,g_p : \mathbb{R}\to\mathbb{R}$ be bounded and Lipschitz continuous, then there exists D_1 independent of m such that*

$$\begin{aligned}&|\mathbb{E}[g_1(X_{i_1})\dots g_k(X_{i_k})g_{k+1}(X_{i_{k+1}})\dots g_p(X_{i_p})]\\ &\qquad - \mathbb{E}[g_1(X_{i_1})\dots g_k(X_{i_k})]\mathbb{E}[g_{k+1}(X_{i_{k+1}})\dots g_p(X_{i_p})]| \leq D_1(a_{\lfloor m/3\rfloor} + \beta_{\lfloor m/3\rfloor}),\end{aligned}$$

where $1 \leq i_1 \leq \ldots \leq i_p \leq T$ and $m = i_{k+1} - i_k$.

The proof is completely analogous to that of Proposition 8.
The next proposition is a Marcinkiewicz-Zygmund type inequality which bounds the $p-$th moment of the sum of the process.

Proposition 10. *Let $(X_t)_{t\in\mathbb{N}}$ be stationary and P-NED on an absolutely regular process $(Z_t)_{t\in\mathbb{Z}}$ with approximation constants $(a_k)_{k\in\mathbb{N}}$, functions $(f_k)_{k\in\mathbb{N}}$, error function Φ and regularity coefficients $(\beta_k)_{k\in\mathbb{N}}$. There furthermore exists $a > p \in \mathbb{N}$ such that $\mathbb{E}(|X_1|^a) < \infty$, $\int_0^1 x^{\frac{a}{a-p+1}-1}\Phi(x)dx < \infty$, $\int_1^\infty x^{a-1}\Phi(x)dx < \infty$ and $\sum_{i=1}^\infty a_i^{\frac{a-p+1}{a}} i^{p-1}$ as well as $\sum_{i=1}^\infty \beta_i^{\frac{a-p}{a}} i^{p-1} < \infty$, then there exist G_1 and G_2 such that*

$$|\mathbb{E}(\sum_{i=1}^T X_i)^p| \leq TG_1|\mathbb{E}(X_1)|^p + G_2T^{\lfloor p/2\rfloor},\ \forall T\in\mathbb{N}. \tag{4.17}$$

Proof of Proposition 10. First notice that

$$|\mathbb{E}(\sum_{i=1}^T X_i)^p| \leq \sum_{i_1,\ldots,i_p=1}^T |\mathbb{E}(X_{i_1}\dots X_{i_p})|. \tag{4.18}$$

We actually show the result by induction applied to the right hand side of (4.18). For $p=1$ the right hand side of (4.18) is obviously bounded by the right hand side of (4.17). For the induction step $p\to p+1$ we want to split the expectations where the time difference is largest. Ordering of time indices yields

$$\sum_{i_1,\ldots,i_{p+1}=1}^T |\mathbb{E}(X_{i_1}\dots X_{i_{p+1}})| = \sum_{1\leq i_1\leq\ldots\leq i_{p+1}}^T |\mathbb{E}(X_{i_1}\dots X_{i_{p+1}})\gamma(i_1,\ldots,i_{p+1})| \tag{4.19}$$

where $\gamma(i_1, \ldots, i_{p+1})$ denotes the number of possible permutations which is smaller or equal to $(p+1)!$. Let $j_s = i_s - i_{s-1},\ s = 2, \ldots, p+1$ and $j_1 = i_1$, then (4.19) is bounded by

$$(p+1)! \sum_{j_1,\ldots,j_{p+1}\geq 0\ j_1+\ldots+j_{p+1}\leq T}^{T} |\mathbb{E}(X_{j_1} X_{j_1+j_2} \ldots X_{j_1+\ldots+j_{p+1}})|. \tag{4.20}$$

We divide the sum (4.20) into $A_1, \ldots, A_p$ where A_s contains all expectations in (4.20) where j_{s+1} is the maximum of $j_2, \ldots, j_{p+1}$[1] and denote I_s the related index set. If j_{s+1} is maximal the other indices can only assume the values $0, \ldots, j_{s+1}$ resulting in less than $(p+1)!T(j_{s+1}+1)^{p-1}$ summands for fixed j_{s+1}. Proposition 7 yields

$$\begin{aligned}
|A_s| &= (p+1)! \sum_{i_1,\ldots,i_{p+1}\in I_s} |\mathbb{E}(X_{i_1} \ldots X_{i_{p+1}})| \\
&\leq (p+1)! \sum_{i_1,\ldots,i_{p+1}\in I_s} |\mathbb{E}(X_{i_1} \ldots X_{i_s})\mathbb{E}(X_{i_{s+1}} \ldots X_{i_{p+1}})| \qquad (4.21)\\
&+ (p+1)!T \sum_{j_s=0}^{T} \left(C_{13} a_{\lfloor j_s/3 \rfloor}^{\frac{a-p}{a}} + C_{14} \beta_{\lfloor j_s/3 \rfloor}^{\frac{a-p-1}{p}} \right) (j_s+1)^{p-1} \qquad (4.22)
\end{aligned}$$

where the sum (4.22) is $O(T)$ by assumption. We use the induction hypothesis for (4.21) to obtain

$$\begin{aligned}
&(p+1)! \sum_{i_1,\ldots,i_s=1}^{T} |\mathbb{E}(X_{i_1} \ldots X_{i_s})| \sum_{i_{s+1},\ldots,i_{p+1}=1}^{T} |\mathbb{E}(X_{i_{s+1}} \ldots X_{i_{p+1}})| \\
&\leq (p+1)!(G_1 T^s |\mathbb{E}(X_1)|^s + G_2 T^{\lfloor \frac{s}{2} \rfloor})(\tilde{G}_1 T^{p+1-s} |\mathbb{E}(X_1)|^{p+1-s} + \tilde{G}_2 T^{\lfloor \frac{p+1-s}{2} \rfloor}) \\
&\leq (p+1)! \Big(G_1 \tilde{G}_1 T^{p+1} |\mathbb{E}(X_1)|^{p+1} + G_2 \tilde{G}_2 T^{\lfloor \frac{p+1}{2} \rfloor} \\
&+ G_1 \tilde{G}_2 T^{\lfloor \frac{p+1-s}{2} \rfloor} T^s |\mathbb{E}(X_1)|^s + \tilde{G}_1 G_2 T^{\lfloor \frac{s}{2} \rfloor} T^{p+1-s} |\mathbb{E}(X_1)|^{p+1-s} \Big) \\
&\leq \hat{G}_1 T^{p+1} |\mathbb{E}(X_1)|^{p+1} + \hat{G}_2 T^{\lfloor \frac{p+1}{2} \rfloor}
\end{aligned}$$

which completes the proof. □

There is also a version for bounded processes.

Proposition 11. *Let $(X_t)_{t\in\mathbb{N}}$ be stationary and P-NED on an absolutely regular process $(Z_t)_{t\in\mathbb{Z}}$ with approximation constants $(a_k)_{k\in\mathbb{N}}$, functions $(f_k)_{k\in\mathbb{N}}$, error function Φ and absolutely regularity coefficients $(\beta_k)_{k\in\mathbb{N}}$. There furthermore exists $D \in \mathbb{R}$ such that $|X_1| < D$ a.s., $\int_0^\infty \Phi(x)dx < \infty$ and $\sum_{i=1}^\infty a_i i^{p-1}$ as well as $\sum_{i=1}^\infty \beta_i i^{p-1} < \infty$, then there exist G_1 and G_2 such that*

$$|\mathbb{E}(\sum_{i=1}^{N} X_i)^p| \leq G_1(T|\mathbb{E}(X_1)|)^p + G_2 T^{\lfloor p/2 \rfloor},\ \forall T \in \mathbb{N}. \tag{4.23}$$

The proof is analogous to that of Proposition 10 using Proposition 8 instead of Proposition 7.

The proof of Theorem 3 consists of four steps:

1. the mean function of $(W_{N,T}(x))_{x\in[0,1]}$ converges to that of $(\Gamma(x))_{x\in[0,1]}$,

[1] To obtain a unique partition we add summands to the sum with the smallest index, if the maximum is attained more than once.

2. the covariance function of $(W_{N,T}(x))_{x\in[0,1]}$ converges to that of $(\Gamma(x))_{x\in[0,1]}$,
3. all finite dimensional distributions of $(W_{N,T}(x))_{x\in[0,1]}$ converge against a multivariate normal distribution,
4. $(W_{N,T}(x))_{x\in[0,1]}$ is tight.

Let without loss of generality $\mathbb{E}(Y_{i,1}) = 0$ and denote $\gamma_i(h) = \mathbb{E}(Y_{i,t}Y_{i,t+h})$ the autocovariance of lag $h \in \mathbb{N}$. Furthermore we want to concentrate on the bounded case $|Y_{i,1}| < c_1,\ i = 1\ldots, N$. The proof in the unbounded case is completely analogous. However, the covariance inequalities then also depend on a, making calculations a little more extensive and harder to understand.

Step 1 of the proof of Theorem 3. First we look at one individual i. Let x, $0 < x < 1$, be arbitrary and $k = \lfloor Tx \rfloor$

$$\begin{aligned}
\mathbb{E}([S_T^{(i)}(x)]^2) &= \frac{1}{Tv_i^2}\mathbb{E}\left(\left[\frac{T-k}{T}\sum_{t=1}^{k} Y_{i,t} - \frac{k}{T}\sum_{t=k+1}^{T} Y_{i,t}\right]^2\right)\\
&= \frac{1}{Tv_i^2}\left(\frac{T-k}{T}\right)^2 \underbrace{\mathbb{E}\left(\left[\sum_{t=1}^{k} Y_{i,t}\right]^2\right)}_{A1}\\
&\quad - \frac{1}{Tv_i^2}2\frac{(T-k)k}{T^2}\underbrace{\mathbb{E}\left(\left[\sum_{t=1}^{k} Y_{i,t}\right]\left[\sum_{t=k+1}^{T} Y_{i,t}\right]\right)}_{A_2} + \frac{1}{Tv_i^2}\frac{k^2}{T^2}\underbrace{\mathbb{E}\left(\left[\sum_{t=k+1}^{T} Y_{i,t}\right]^2\right)}_{A3}.
\end{aligned}$$

Elementary calculations yield

$$\begin{aligned}
A_1 &= kv_i^2 - 2k\sum_{h=k}^{\infty}\gamma_i(h) - 2\sum_{h=1}^{k-1} h\gamma_i(h)\\
A_2 &= \sum_{h=1}^{T}\gamma_i(h)\min(h, k, T-k, T-h)\\
A_3 &= (T-k)v_i^2 - 2(T-k)\sum_{h=T-k}^{\infty}\gamma_i(h) - 2\sum_{h=1}^{T-k-1} h\gamma_i(h)
\end{aligned}$$

and therefore

$$\begin{aligned}
\mathbb{E}(W_{N,T}(x)) = \frac{\sqrt{N}}{T}\frac{1}{N}\sum_{i=1}^{N}\Bigg(&\frac{1}{v_i^2}\left(\frac{T-k}{T}\right)^2\left(-2k\sum_{h=k}^{\infty}\gamma_i(h) - 2\sum_{h=1}^{k-1} h\gamma_i(h)\right)\\
&+ \frac{1}{v_i^2}\frac{(T-k)k}{T^2}\sum_{h=1}^{T}\gamma_i(h)\min(h, k, T-k, T-h)\\
&+ \frac{1}{v_i^2}\frac{k^2}{T^2}\left(-2(T-k)\sum_{h=T-k}^{\infty}\gamma_i(h) - 2\sum_{h=1}^{T-k-1} h\gamma_i(h)\right)\Bigg).
\end{aligned}$$

Using Assumption 1 2) a) ii) one has:

$$|\mathbb{E}(W_{N,T}(x))| \le \frac{\sqrt{N}}{T}\frac{1}{N}\sum_{i=1}^{N}\frac{C_{15}}{\delta^2}\left(5\underbrace{\sum_{h=1}^{\infty} h(1+h)^{-b}}_{B_1} + 2\underbrace{k\sum_{h=k}^{\infty}(1+h)^{-b}}_{B_2} + 2\underbrace{(T-k)\sum_{h=T-k}^{\infty}(1+h)^{-b}}_{B_3}\right.$$

where by the integral criteria for sums

$$B_1 \leq \int_0^\infty (1+h)^{-b+1} dh = \frac{1}{b-2} \text{ and } B_2 \leq k \int_{k-1}^\infty c(1+h)^{-b} dh = \frac{k^{-b+2}}{b-1} \leq \frac{1}{b-1}.$$

and analogously $B_3 \leq \frac{(T-k)^{-b+2}}{b-1}$. By Assumption i) of Theorem 3 $|\mathbb{E}(W_{N,T}(x))| \to 0$ which gives the desired mean structure. □

Step 2 of the proof of Theorem 3. Calculating the covariance structure is more tedious. By Assumption 1 I) we have

$$\text{Cov}(W_{N,T}(x), W_{N,T}(y)) = \frac{1}{N} \sum_{i=1}^N Cov([S^{(i)}(x)]^2, [S^{(i)}(y)]^2),$$

so it is enough to compute the covariance structure of one individual i, $i = 1, \ldots, N$. We denote therefore $k = [xn] < [yn] = j$ and expand $E([S^{(i)}(x)]^2 [S^{(i)}(y)]^2)$ to obtain

$$\begin{aligned}
E([S^{(i)}(x)]^2[S^{(i)}(y)]^2) = & \frac{(T-k)^2(T-j)^2}{T^6 v_i^4} \mathbb{E}\left(\sum_{s,t,u,v=1}^{k,k,j,j} Y_{i,t} Y_{i,s} Y_{i,u} Y_{i,v} \right) \\
& - 2\frac{(T-k)^2(T-j)j}{T^6 v_i^4} \mathbb{E}\left(\sum_{s,t,u=1,v=j+1}^{k,k,j,T} Y_{i,s} Y_{i,t} Y_{i,u} Y_{i,v} \right) \\
& + \frac{(T-k)^2 j^2}{T^6 v_i^4} \mathbb{E}\left(\sum_{s,t=1,u,v=j+1}^{k,k,T,T} Y_{i,s} Y_{i,t} Y_{i,u} Y_{i,v} \right) \\
& - 2\frac{(T-k)k(T-j)^2}{T^6 v_i^4} \mathbb{E}\left(\sum_{s=1,t=k+1,u,v=1}^{k,T,j,j} Y_{i,s} Y_{i,t} Y_{i,u} Y_{i,v} \right) \\
& + 4\frac{(T-k)k(T-j)j}{T^6 v_i^4} \mathbb{E}\left(\sum_{s=1,t=k+1,u=1,v=j+1}^{k,T,j,T} Y_{i,s} Y_{i,t} Y_{i,u} Y_{i,v} \right) \\
& - 2\frac{(T-k)kj^2}{T^6 v_i^4} \mathbb{E}\left(\sum_{s=1,t=k+1,u,v=j+1}^{k,T,T,T} Y_{i,s} Y_{i,t} Y_{i,u} Y_{i,v} \right) \\
& + \frac{k^2(T-j)^2}{T^6 v_i^4} \mathbb{E}\left(\sum_{s,t=k+1,u,v=1}^{T,T,j,j} Y_{i,s} Y_{i,t} Y_{i,u} Y_{i,v} \right) \\
& - 2\frac{k^2(T-j)j}{T^6 v_i^4} \mathbb{E}\left(\sum_{s,t=k+1,u=1,v=j+1}^{T,T,j,T} Y_{i,s} Y_{i,t} Y_{i,u} Y_{i,v} \right) \\
& + \frac{k^2 j^2}{T^6 v_i^4} \mathbb{E}\left(\sum_{s,t=k+1,u,v=j+1}^{T,T,T,T} Y_{i,s} Y_{i,t} Y_{i,u} Y_{i,v} \right) \\
& = A_{i,1} + A_{i,2} + A_{i,3} + A_{i,4} + A_{i,5} + A_{i,6} + A_{i,7} + A_{i,8} + A_{i,9}.
\end{aligned}$$

Exemplarily we look at $A_{i,1}$ which we divide into

$$\begin{aligned}A_{i,1} &= \frac{(T-k)^2(T-j)^2}{T^6 v_i^4}\left[\mathbb{E}\left(\sum_{s,t,u,v=1}^{k,k,k,k} Y_{i,t}Y_{i,s}Y_{i,u}Y_{i,v}\right) + \mathbb{E}\left(\sum_{s,t=1,u,v=k+1}^{k,k,j,j} Y_{i,t}Y_{i,s}Y_{i,u}Y_{i,v}\right)\right]\\ &= \frac{(T-k)^2(T-j)^2}{T^6 v_i^4}(B_{i,1}+B_{i,2}).\end{aligned}$$

We split $B_{1,i}$ further in parts which are substantial and parts which are negligible. Ordering such that $t \leq s \leq u \leq v$ and substituting $s-t=l,\ u-s=m,\ v-u=n$ yields

$$B_{1,i} = \sum_{t+l+m+n\leq k} \mathbb{E}\left(Y_{i,t}Y_{i,t+s}Y_{i,t+s+m}Y_{i,t+s+m+n}\right) a(s,m,n)$$

where $a(s,m,n)$ equals the number of permutations of $Y_{i,t}Y_{i,s+t}Y_{i,t+s+m}Y_{i,t+s+m+n}$. Now we want to apply Proposition 8 and split the expectations where the lag difference between the random variables is largest:

$$\begin{aligned}B_{1,i} &= \sum_{t+s+m+n\leq k,\ m,n<s} \mathbb{E}(Y_{i,t})\mathbb{E}(Y_{i,t+s}Y_{i,t+s+m}Y_{i,t+s+m+n})a(s,m,n) + R_{1,i}\\ &+ \sum_{t+s+m+n\leq k,\ s,n\leq m} \mathbb{E}(Y_{i,t}Y_{i,t+s})\mathbb{E}(Y_{i,t+s+m}Y_{i,t+s+m+n})a(s,m,n) + R_{2,i}\\ &+ \sum_{t+s+m+n\leq k,\ s,m<n} \mathbb{E}(Y_{i,t}Y_{i,t+s}Y_{i,t+s+m})\mathbb{E}(Y_{i,t+s+m+n})a(s,m,n) + R_{3,i}\\ &= R_{1,i}+R_{2,i}+R_{3,i}\\ &+ \sum_{t+s+m+n\leq k,\ s,n\leq m} \mathbb{E}(Y_{i,t}Y_{i,t+s})\mathbb{E}(Y_{i,t+s+m}Y_{i,t+s+m+n})a(s,m,n)\end{aligned}$$

where by Proposition 8 and the integral criteria

$$\begin{aligned}|R_{1,i}+R_{2,i}+R_{3,i}| &\leq 3\cdot 24 C_{16} \sum_{t+s+m+n\leq k,\ m,n\leq s} (1+s)^{-b}\\ &\leq 72kC_{16}\sum_{s=0}^{\infty}(s+1)^2 c(1+s)^{-b} \leq \frac{kC_{17}}{b-3}\end{aligned}$$

and therefore $\sup_{i=1,\dots,N} \frac{(T-k)^2(T-j)^2}{T^6 v_i^4}|R_{1,i}+R_{2,i}+R_{3,i}| \to 0$. Now we turn towards the non vanishing part of $B_{1,i}$. We change summation another time to arrive at

$$\begin{aligned}&\sum_{t+s+m+n\leq k,\ s,n\leq m} \mathbb{E}(Y_{i,t}Y_{i,t+s})\mathbb{E}(Y_{i,t+s+m}Y_{i,t+s+m+n})a(l,m,n)\\ &= \sum_{s=0}^{k}\sum_{n=0}^{k} 3 \sum_{m=\max(s,n)}^{k-s-n} (k-s-n-\max(s,n)-m)\gamma_i(s)\gamma_i(n)b(s,n)\\ &= 3\sum_{s=0}^{k}\sum_{n=0}^{k}\gamma_i(s)\gamma_i(n)b(s,n)\left((k-s-n-\max(s,n)+1)(k-s-n-\max(s,n))\right.\\ &\left.-\frac{(k-s-n+1)(k-s-n)}{2}+\frac{\max(s,n)(\max(s,n)-1)}{2}\right) = \frac{3}{2}k^2v_i^4 + 3U_i \qquad (4.24)\end{aligned}$$

where $b(i,j) = \begin{cases} 1 & i,j=0 \\ 2 \quad \text{for} & (i,j)=(0,1)\wedge(1,0) \\ 4 & i,j>0 \end{cases}$

and the factor 3 appears since there are three possibilities to partition four random variables into two pairs. The error U_i consists of a sum of autocovariances which do not appear in (4.24) and one of autocovariances of improper quantity

$$\begin{aligned}
|U_i| &\leq 12k^2 \Big| \sum_{m=0}^{\infty} \sum_{n=k+1}^{\infty} \gamma_i(m)\gamma_i(n) \Big| + 12 \Big| \sum_{s=0}^{k} \sum_{n=0}^{k} \gamma(s)\gamma(n) \left[5k(s+n) + (s+n)^2\right] \Big| \\
&\leq C_{18} k^2 \int_k^{\infty} (1+s)^{-b} ds + C_{19} k \left(\int_0^{\infty} (1+s)^{-b+1} ds \right)^2 + C_{20} \left(\int_0^{\infty} (1+s)^{-b+2} ds \right)^2 \\
&= C_{18} \frac{k^{-b+3}}{b-1} + C_{19} k \left(\frac{1}{b-2} \right)^2 + C_{20} \left(\frac{1}{b-3} \right)^2
\end{aligned}$$

where the squared integrals arise through the integral criteria applied to the separated sums. Therefore we get $\sup_{i=1,\ldots,N} \frac{(T-k)^2(T-j)^2}{T^6 v_i^4} |U_i| \to 0$. By analogous calculations one gets

$$\begin{aligned}
A_{1,1} &= \frac{(T-k)^2(T-j)^2(3k^2 + k(j-k))}{2T^6} + R_{i,4} \\
A_{i,2} &= R_{i,5} \\
A_{i,3} &= \frac{(T-k)^2 j^2 k(T-j)}{2T^6} + R_{i,6} \\
A_{i,4} &= -2\frac{(T-k)k(T-j)^2 2k(j-k)}{2T^6} + R_{i,7} \\
A_{i,5} &= 4\frac{(T-k)k(T-j)^2 jk}{2T^6} + R_{i,8} \\
A_{i,6} &= R_{i,9} \\
A_{i,7} &= \frac{k^2(T-j)^2((T-j)(j-k) + k(j-k) + (T-j)k + 3(j-k)^2)}{2T^6} + R_{i,10} \\
A_{i,8} &= -2\frac{k^2(T-j)^2 j 2(j-k)}{2T^6} + R_{i,11} \\
A_{i,9} &= \frac{k^2 j^2 (3(T-j)^2 + (T-j)(j-k))}{2T^6} + R_{i,12}
\end{aligned}$$

with $\sup_{i \in 1,\ldots,N} \sum_{k=4}^{12} |R_{i,k}| \to 0$. Term manipulations yield the desired covariance structure. □

Step 3 of the Proof of Theorem 3. We show convergence of finite dimensional distributions. Let $k \in \mathbb{N}$ and $x_1, \ldots, x_k$ be arbitrary, by the Cramer Wold device it is sufficient (and necessary) to show convergence of linear combinations for arbitrary $\lambda_1, \ldots, \lambda_k$. By change of summation we get

$$\sum_{j=1}^{k} \lambda_j W_{N,T}(x_j) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \underbrace{\sum_{j=1}^{k} \lambda_j \left(\left(S_T^{(i)}(x_j) \right)^2 - \frac{\lfloor x_j T \rfloor (T - \lfloor x_j T \rfloor)}{T^2} \right)}_{D_i}$$

where D_i are independent but not identically distributed. Since D_i also depends on T, we apply a central limit theorem for random arrays of Lyapunov type (see for example Serfling (1980) p. 30). Therefore one needs to show

$$\frac{\sum_{i=1}^{N} \mathbb{E} \left(\sum_{j=1}^{k} \lambda_j [S_i^2(x_j) - \mathbb{E}\{S_i^2(x_j)\}] \right)^4}{\left(\sum_{i=1}^{N} \mathbb{E} \left[\sum_{j=1}^{k} \lambda_j \left\{ S_i^2(x_j) - \mathbb{E}(S_i^2(x_j)) \right\} \right]^2 \right)^2} \to 0. \tag{4.25}$$

For the nominator we repeatedly apply the c_r inequality and Proposition 11 to get

$$\begin{aligned}
&\sum_{i=1}^{N}\mathbb{E}\left(\sum_{j=1}^{k}\lambda_j[S_i^2(x_j)-\mathbb{E}\{S_i^2(x_j)\}]\right)^4 \\
&\quad\le 8^k\sum_{j=1}^{k}\lambda_j^4\sum_{i=1}^{N}\mathbb{E}(S_i^2(x_j)-\mathbb{E}[S_i^2\{x_j\}])^4 \\
&\quad\le 8^{k+1}\sum_{j=1}^{k}\lambda_j^4\sum_{i=1}^{N}\mathbb{E}(S_i^8(x_j))+\mathbb{E}(S_i^2(x_j))^4) \\
&\quad\le 8^{k+2}\sum_{j=1}^{k}\lambda_j^4\sum_{i=1}^{N}\left(\frac{1}{T^4}\mathbb{E}\left[\sum_{t=1}^{\lfloor x_jT\rfloor}Y_{i,t}\right]^8+\frac{1}{T^4}\mathbb{E}\left[\sum_{t=\lfloor x_jT\rfloor+1}^{T}Y_{i,t}\right]^8\right. \\
&\quad\left.+\frac{1}{T^4}\left[\mathbb{E}\left(\sum_{t=1}^{\lfloor x_jT\rfloor}Y_{i,t}\right)^2\right]^4+\frac{1}{T^4}\left[\mathbb{E}\left(\sum_{t=\lfloor x_jT\rfloor+1}^{T}Y_{i,t}\right)^2\right]^4\right) \\
&\quad\le 8^{k+2}\sum_{j=1}^{k}\lambda_j^4N(G_2+\tilde{G}_2^4).
\end{aligned}$$

For the denominator we exploit the cross-sectional independence and arrive at

$$\begin{aligned}
\left(\sum_{i=1}^{N}\mathbb{E}\left(\sum_{j=1}^{k}\lambda_j(S_i^2(x_j)-\mathbb{E}(S_i^2(x_j)))\right)^2\right)^2 &= \left(\sum_{i=1}^{N}\sum_{j,l=1}^{k}\lambda_j\lambda_l\mathrm{Cov}(S_i^2(x_j),S_i^2(x_l))\right)^2 \\
&= \left(\sum_{i=1}^{N}\underbrace{\sum_{j,l=1}^{k}\lambda_j,\lambda_l\mathrm{Cov}(\Gamma(x_j),\Gamma(x_l))+R_i}_{M}\right)^2 \\
&= M^2N^2+2MN\sum_{i=1}^{N}R_i+(\sum_{i=1}^{N}R_i)^2
\end{aligned}$$

where R_i denotes the remainder fulfilling $\sup_{i=1,\dots,N}R_i\to 0$, see the second step of the proof of Theorem 1. Since the Gaussian process Γ possesses a positive definite covariance function, we have $M>0$ and the denominator grows of the order N^2 while the nominator only grows linearly in N, which proofs (4.25) and hence the asymptotic normality. Together with step 1 and step 2 this proves that the finite dimensional distributions of $(W_{N,T}(x))_{x\in[0,1]}$ converge against that of $(\Gamma(x))_{x\in[0,1]}$. □

Step 4 of the Proof of Theorem 3. We want to apply the moment criteria of Billingsley (1968) (see page 95) and therefore look at the difference between $W_{N,T}(x)$ and $W_{N,T}(y)$. By an expansion we see that we need bounds for four different moments

$$\begin{aligned}
\mathbb{E}([W_{N,T}(x)-W_{N,T}(y)]^4) \\
&= \frac{1}{N^2}\mathbb{E}\left(\left[\sum_{i=1}^{N}\underbrace{S_i^2(x)-\frac{\lfloor xT\rfloor(T-\lfloor xT\rfloor)}{T^2}-S_i^2(x)+\frac{\lfloor xT\rfloor(T-\lfloor xT\rfloor)}{T^2}}_{M_i}\right]^4\right)\\
&= \frac{1}{N^2}\sum_{i\neq j\neq k\neq l}\mathbb{E}(M_i)\mathbb{E}(M_j)\mathbb{E}(M_k)\mathbb{E}(M_l)+\frac{1}{N^2}\sum_{i\neq j\neq k}\mathbb{E}(M_i^2)\mathbb{E}(M_j)\mathbb{E}(M_k)\\
&+\frac{1}{N^2}\sum_{i\neq j}\mathbb{E}(M_i^3)\mathbb{E}(M_j)+\frac{1}{N^2}\sum_{i\neq j}\mathbb{E}(M_i^2)\mathbb{E}(M_j^2)+\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}(M_i^4).
\end{aligned}$$

In the first step of the proof of Theorem 3 it it shown that

$$|\mathbb{E}(M_i)|\leq\left|\mathbb{E}\left(S_i^2(x)-\frac{\lfloor xT\rfloor(T-\lfloor xT\rfloor)}{T^2}\right)\right|+\left|\mathbb{E}\left(S_i^2(y)+\frac{\lfloor yT\rfloor(T-\lfloor yT\rfloor)}{T^2}\right)\right|\leq\frac{C_{21}}{T},$$

where C_{21} is independent of i, x and y. Let $\lfloor Tx\rfloor = u < \lfloor Ty\rfloor = v$, for $\mathbb{E}(M_i^2)$ we apply Cauchy-Schwarz and c_r inequality to get

$$\begin{aligned}
\mathbb{E}(M_i^2) &\leq 2\mathbb{E}([S_i(x)^2-S_i(y)^2]^2)+2\left(\frac{\lfloor xT\rfloor(T-\lfloor xT\rfloor)}{T^2}-\frac{\lfloor yT\rfloor(T-\lfloor yT\rfloor)}{T^2}\right)^2\\
&\leq\sqrt{\mathbb{E}([S_i(x)-S_i(y)]^4)}\sqrt{8\mathbb{E}([S_i(x)]^4)+8\mathbb{E}([S_i(y)]^4)}+2\left(\frac{(v-u)(u-T-v)}{T^2}\right)^2.
\end{aligned}$$

Using Proposition 11 one has

$$|\mathbb{E}(S_i(x)^4)|\leq\frac{8}{T^2\delta^4}\mathbb{E}\left(\frac{T-u}{T}\sum_{t=1}^{u}Y_{i,t}\right)^4+\frac{8}{T^2\delta^4}\mathbb{E}\left(\frac{u}{T}\sum_{t=u+1}^{T}Y_{i,t}\right)^4\leq C_{22}$$

and also

$$\mathbb{E}([S_i(x)-S_i(y)]^4)=\frac{8}{T^2\delta^4}\mathbb{E}\left(\left[\sum_{t=u+1}^{v}Y_{i,t}\right]^4\delta^4\right)+\frac{8(v-u)^4}{T^6}\mathbb{E}\left(\sum_{t=1}^{T}Y_{i,t}\right)^4\leq C_{23}\frac{(u-v)^2}{T^2}.$$

Together we have $\mathbb{E}(M_i^2)\leq C_{24}\frac{v-u}{T}$ and analogously $\mathbb{E}(M_i^4)\leq C_{25}\frac{(v-u)^2}{T^2}$ respectively $|\mathbb{E}(M_i^3)|\leq C_{26}\frac{(v-u)^{3/2}}{T^{3/2}}$. Since $N/T\to 0$, there is C_{27} such that $N/T\leq C_{27}$[2] and we arrive at

$$\begin{aligned}
\mathbb{E}([W_{N,T}(x)-W_{N,T}(y)]^4) &\leq\frac{1}{N^2}N^4\left(\frac{C_{21}}{T}\right)^4+\frac{1}{N^2}N^3C_{24}\frac{v-u}{T}\left(\frac{C_{21}}{T}\right)^2\\
&+\frac{1}{N^2}N^2C_{26}\frac{(v-u)^{3/2}}{T^{3/2}}\frac{C_{21}}{T}+\left(C_{24}\frac{v-u}{T}\right)^2+\frac{C_{25}}{N}\frac{(v-u)^2}{T^2}\\
&\leq C_{28}\frac{(v-u)^2}{T^2}+\frac{C_{29}}{T^2}+\frac{C_{30}(v-u)}{T^2}+\frac{C_{31}(v-u)^{3/2}}{T^{5/2}}\leq C_{32}|x-y|^2
\end{aligned}$$

which proves tightness of the process. □

[2] To emphasise that N,T jointly tend to infinity it is maybe more convenient to use $N(T)$ instead of N (respectively $T(N)$ instead of T). We have forgone on it for a better readability. However, the more elaborate notation is here superior. Since what is meant and used is that $N(T)/T\leq C_{27}$, $\forall T\in\mathbb{N}$.

Proof of Theorem 4. We assume that $|Y_{i,1}| \leq c_1,\ i = 1, \dots, N$, though the proof is analogous in the unbounded case. We first calculate the mean squared error of $\hat{v}_i$. Equations for variance and bias of $\hat{v}_i$ are already known, see for example Anderson (1971) chapters eight and nine, though these are for known location. Denote therefore

$$\tilde{v}_i = \tilde{\gamma}_{i,0} + 2\sum_{h=1}^{b_T} \tilde{\gamma}_{i,h} k\left(\frac{h}{b_T}\right) \quad \text{with} \quad \tilde{\gamma}_i(h) = \frac{1}{T}\sum_{t=1}^{N-h} \left(Y_{i,t} - \mathbb{E}(Y_{i,1})\right)\left(Y_{i,t+h} - \mathbb{E}(Y_{i,1})\right)$$

the long run variance with known location, where we continue to assume that $\mathbb{E}(Y_{i,1}) = 0,\ i = 1, \dots, N$. We denote $\overline{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{i,t}$ and describe $\hat{v}_i$ by $\tilde{v}_i$:

$$\begin{aligned}
\hat{v_i}^2 &= \tilde{v}_i^2 - \frac{1}{T}\sum_{t=1}^{T} Y_{i,t}\overline{Y}_i - 2\sum_{h=1}^{b_T}\frac{1}{T}\sum_{t=1}^{T-h} Y_{i,t+h}\overline{Y}_i k\left(\frac{h}{b_T}\right) \\
&\quad - \frac{1}{T}\sum_{t=1}^{T} Y_{i,t}\overline{Y}_i - 2\sum_{h=1}^{b_T}\frac{1}{T}\sum_{t=1}^{T-h} Y_{i,t}\overline{Y}_i k\left(\frac{h}{b_T}\right) + \sum_{h=-b_T}^{b_T}\frac{1}{T}\overline{Y_i}^2 k\left(\frac{h}{b_T}\right) \\
&= \tilde{v_i}^2 + R_1 + R_2 + R_3 + R_4 + R_5.
\end{aligned}$$

Following the calculations of step 2 of the proof of Theorem 3 one finds upper bounds for the errors $R_1 + R_2$ respectively $R_3 + R_4$

$$E([R_1 + R_2]^2) \leq \frac{1}{T^4} 6 \sum_{h=0}^{b_T}\sum_{k=-b_T}^{b_T} (T^2 v_i^2 + C_{33} T) k\left(\frac{h}{b_T}\right) k\left(\frac{k}{b_T}\right) \leq C_{34}\frac{b_T^2}{T^2}$$

and Proposition 11 yields

$$E(R_3^2) \leq \frac{1}{T^6}\sum_{h=-b_T}^{b_T}\sum_{k=-b_T}^{b_T} T^2 C_{35} k\left(\frac{h}{b_T}\right) k\left(\frac{k}{b_T}\right) \leq \frac{b_T^2}{T^4} C_{36}.$$

Therefore we get

$$\mathbb{E}\left([\hat{v}_i^2 - \tilde{v}_i^2]^2\right) \leq C_{37}\frac{b_T^2}{T^2}. \tag{4.26}$$

In the next step we calculate the bias of $\tilde{v}_i^2$. straightforward calculations yield:

$$\begin{aligned}
|E(\tilde{v}_i^2) - v_i^2| &= \left| E\left(\frac{1}{T}\sum_{t=1}^{T} Y_{i,t}^2 + 2\frac{1}{T}\sum_{h=1}^{b_T}\sum_{t=1}^{T-h} Y_{i,t}Y_{i,t+h} k\left(\frac{h}{b_T}\right)\right) - \sum_{h=-\infty}^{\infty}\gamma_i(h)\right| \\
&\leq \left| 2\sum_{h=1}^{b_T}\left(\left(1 - \frac{h}{T}\right) k\left(\frac{h}{b_T}\right) - 1\right)\gamma_i(h)\right| + 2\left|\sum_{h=b_T+1}^{\infty}\gamma_i(h)\right| \\
&\leq 2\left|\sum_{h=1}^{b_T}\left(1 - k\left(\frac{h}{b_T}\right)\right)\gamma_i(h)\right| + 2\left|\sum_{h=1}^{b_T}\gamma_i(h) k\left(\frac{h}{b_T}\right)\frac{h}{T}\right| + 2\left|\sum_{h=b_T+1}^{\infty}\gamma_i(h)\right| \\
&= A_1 + A_2 + A_3.
\end{aligned}$$

The sum A_1 describes the error which is generated by the kernel. To bound the error, we develop the kernel k around 0. Since the first $m-1$ derivatives are 0 by Assumption 2 iv), the first non-vanishing term is of order m. However to ensure that the remainder is negligible, we only develop the Taylor series up to order $s-1$:

$$|A_1| \leq \sum_{h=1}^{b_T} C_{38}(1+h)^{-b}\left(\frac{h}{b_T}\right)^s \leq \frac{1}{b_T^s} C_{39}\sum_{h=1}^{\infty}(1+h)^{-b+s} = \frac{1}{b_T^s} C_{40}. \tag{4.27}$$

Assumption 1 2) a) ii) and the integral criterion yields $|A_2| \leq \frac{C_{41}}{T}$ and $|A_3| \leq C_{42} b_T^{-b+1}$ and therefore

$$|E(\tilde{v}_i^2) - v_i^2| \leq \frac{1}{b_T^s} C_{40} + \frac{C_{41}}{T} + C_{42} b_T^{-b+1}. \tag{4.28}$$

Now we turn our attention to the variance of $\tilde{v}_i$, where the following expansion is known (see for example Anderson (1971), page 528, chapter 9.3.3)

$$\begin{aligned} \mathrm{Var}(\tilde{v}_i) &= \frac{1}{T} \sum_{g,h=-b_T}^{b_T} k\left(\frac{h}{b_T}\right) k\left(\frac{g}{b_T}\right) \sum_{r=-T+1}^{T-1} \phi(r,g,h) \\ &\cdot (\gamma_i(r)\gamma_i(r+h-g) + \gamma_i(r-g)\gamma_i(r+g) + \kappa_i(h,-r,g-r)) \\ &= B_1 + B_2 + B_3 \end{aligned}$$

where $\kappa_i(r,s,t) = \mathbb{E}(Y_{i,1}Y_{i,r+1}Y_{i,s+1}, Y_{i,t+1}) - \gamma(r)\gamma(t-r) - \gamma(s)\gamma(t-r) - \gamma(t)\gamma(s-r)$ for $r,s,t \in \mathbb{N}$ is the fourth order cumulant and the formal definition of $\phi(r,g,h)$ can be found on page 528 of Anderson (1971). For us it is only important that $|\phi(r,g,h)| \leq 1$. Following the calculations in Anderson (1971) we get

$$\begin{aligned} |B_2| &\leq \left| \frac{1}{T} \sum_{g,h=-b_T}^{b_T} \sum_{r=\max(g-b_T,h-b_T,-T+1)}^{\min(g+b_T,h+b_T,T-1)} \phi(r,r-g,h-r) k\left(\frac{r-g}{b_T}\right) k\left(\frac{h-r}{b_T}\right) \gamma_i(g)\gamma_i(h) \right| \\ &+ \left(\frac{8b_T}{T} + \frac{4}{b_T T}\right) \sum_{g=-\infty}^{\infty} \sum_{h=b_T+1}^{\infty} R|\gamma_i(g)\gamma_i(h)| = D_1 + D_2 \end{aligned}$$

where Proposition 8 yields $|D_1| \leq C_{43}\frac{b_T}{T}$, $|D_2| \leq C_{44}\frac{b_T}{T}$ and similar arguments reveal $|B_1| \leq C_{45}\frac{b_T}{T}$. We rearrange the sum of cumulants and split the expectation with Proposition 8 where the lag difference is largest

$$\begin{aligned} |B_3| &\leq \frac{8R}{T} \sum_{r,s,t=0}^{\infty} |\kappa(r,s,t)| \\ &\leq \frac{24R}{T} \sum_{r,s,t=0}^{\infty} |\kappa(r,r+s,r+s+t)| \\ &\leq \frac{24R}{T} \sum_{r=0}^{\infty} \sum_{s,t \leq r} |\mathbb{E}(Y_{i,1})\mathbb{E}(Y_{i,1+r}Y_{i,1+r+s}Y_{i,1+r+s+t}) - \gamma_i(r)\gamma_i(t)| \\ &+ \frac{24R}{T} \sum_{s=0}^{\infty} \sum_{r,t \leq s} |\mathbb{E}(Y_{i,1}Y_{i,r+1})\mathbb{E}(Y_{i,1+r+s}Y_{i,1+r+s+t}) - \gamma_i(r)\gamma_i(t)| \\ &+ \frac{24R}{T} \sum_{t=0}^{\infty} \sum_{r,s \leq t} |\mathbb{E}(Y_{i,1}Y_{i,r+1}Y_{i,1+r+s})\mathbb{E}(Y_{i,1+r+s+t}) - \gamma_i(r)\gamma_i(t)| \\ &+ \frac{24R}{T} \sum_{r,s,t=0}^{\infty} |\gamma_i(r+s)\gamma_i(s+t)| + \frac{24R}{T} \sum_{r,s,t=0}^{\infty} |\gamma_i(r+s+t)\gamma_i(s)| \\ &= F_1 + F_2 + F_3 + F_4 + F_5. \end{aligned}$$

We look exemplarily at F_1, F_2 and F_4:

$$\begin{aligned}
F_1 &\leq \frac{C_{46}}{T}\sum_{r=0}^{\infty} r^2(1+r)^{-b} + \frac{C_{47}}{T}\sum_{r=0}^{\infty} r(1+r)^{-b} \leq \frac{C_{48}}{T}\\
F_2 &\leq \frac{C_{49}}{T}\sum_{s=0}^{\infty} r^2(1+s)^{-b} + \frac{C_{50}}{T}\sum_{s=0}^{\infty}\sum_{r=s+1}^{\infty}(1+r)^{-b}\sum_{t=0}^{\infty}(1+t)^{-b}\\
&\leq \frac{C_{51}}{T} + \sum_{s=0}^{\infty}\frac{C_{52}}{T}(1+s)^{-b+1} \leq \frac{C_{53}}{T}\\
F_4 &= \frac{C_{54}}{T}\sum_{s=0}^{\infty}\sum_{r=s+1}^{\infty}(1+r)^{-b+1}\sum_{t=s+1}^{\infty}(1+t)^{-b+1} \leq \frac{C_{55}}{T}
\end{aligned}$$

So we finally arrive at

$$\mathrm{Var}(\tilde{v}_i) \leq \frac{C_{56}b_T}{T} \tag{4.29}$$

and therefore by (4.26), (4.28) and (4.29)

$$\mathbb{E}([\hat{v}_i^2 - v_i^2]^2) \leq C_{57}\frac{b_T^2}{T^2} + C_{58}b_T^{-2s} + C_{59}b_T^{-2b+2} + C_{60}\frac{b_T}{T}. \tag{4.30}$$

Now we show that the long run variance estimations $\hat{v}_i$ are bounded from below for large T, N. Denote by $D_{N,T}$ the event that $\hat{v}_i^2 > v_i^2/2,\ i = 1, \ldots, N$, then

$$\begin{aligned}
P(D_{N,T}) &\geq 1 - \sum_{i=1}^{N} P(\hat{v}_i^2 < v_i^2/2)\\
&\geq 1 - \sum_{i=1}^{N} P(|\hat{v}_i^2 - v_i^2| \geq v_i^2/2)\\
&\geq 1 - 4\left(C_{57}\frac{b_T^2}{T^2} + C_{58}b_T^{-2s} + C_{59}b_T^{-2b+2} + C_{60}\frac{b_T}{T}\right)\sum_{i=1}^{N}\frac{1}{v_i^4} \to 1
\end{aligned}$$

and therefore it is enough to prove Theorem 2 on $D_{N,T}$. In the following we show that the difference between $W_{N,T}(x)$ and $\tilde{W}_{N,T}(x)$ is negligible for every $x \in [0,1]$. By an expansion one can extract the error of the long run variance estimation from this difference

$$\begin{aligned}
&W_{N,T}(x) - \tilde{W}_{N.T}(x)\\
&= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{1}{\hat{v}_i^2} - \frac{1}{v_i^2}\right)\left(\frac{1}{T}\left[\sum_{t=1}^{\lfloor Tx\rfloor} Y_{i,t} - \frac{\lfloor Tx\rfloor}{T}\sum_{t=1}^{T} Y_{i,t}\right]^2 - v_i^2\frac{\lfloor Tx\rfloor(T-\lfloor Tx\rfloor)}{T^2}\right)\\
&+ \frac{\lfloor Tx\rfloor(T-\lfloor Tx\rfloor)}{T^2}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{v_i^2 - \hat{v}_i^2}{\hat{v_i}^2} = G_1 + G_2.
\end{aligned} \tag{4.31}$$

The second term of (4.31) can be bounded by (4.26), (4.28) and (4.30)

$$\begin{aligned}
\mathbb{E}\left(G_2^2\right) &\leq \frac{1}{N}\sum_{i\neq j}\frac{4}{v_i^2v_j^2}|\mathbb{E}(\hat{v}_i^2 - v_i^2)\mathbb{E}(\hat{v}_j^2 - v_j^2)| + \frac{1}{N}\sum_{i=1}^{N}\frac{4}{v_i^4}\mathbb{E}(\hat{v}_i^2 - v_i^2)^2\\
&\leq C_{61}N\left(\frac{b_T^2}{T^2} + \frac{1}{b_T^{2s}} + \frac{1}{T^2} + b_T^{-2b+2}\right) + C_{62}\left(\frac{b_T^2}{T^2} + b_T^{-2s} + b_T^{-2b+2} + \frac{b_T}{T}\right) \to 0.
\end{aligned}$$

Using the Cauchy-Schwarz inequality and (4.30) one gets for E_1 :

$$
\begin{aligned}
\mathbb{E}(|E_1|) \leq & \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{4}{\delta^4} \sqrt{\mathbb{E}([\hat{v}_i{}^2 - v_i^2]^2)} \\
& \sqrt{\mathbb{E}\left(\left[\frac{1}{T}\left\{\sum_{t=1}^{\lfloor Tx \rfloor} Y_{i,t} - \frac{\lfloor Tx \rfloor}{T} \sum_{t=1}^{T} Y_{i,t}\right\}^2 - v_i^2 \frac{\lfloor Tx \rfloor (T - \lfloor Tx \rfloor)}{T^2}\right]^2\right)} \\
& \leq \sqrt{N} \sqrt{C_{62}\left(\frac{b_T^2}{T^2} + b_T^{-2s} + b_T^{-2b+2} + \frac{b_T}{T}\right)} \sqrt{\left(1 + \frac{C_{63}}{T^2}\right)} \to 0
\end{aligned}
$$

where C_{63} can be calculated based on step 1 and 2 of the Proof of Theorem 1. The tightness of $(\tilde{W}_{N,T}(x) - W_{N,T}(x))_{x \in [0,1]}$ can be proved like the tightness of $W_{N,T}(x)$ using that $\hat{v}_i$ is bounded from below, which completes the proof. □

The proof of Theorem 5 consists of three steps:

1. showing pointwise convergence $\check{W}_{N,T}(x) - \check{W}_{N,T}(x) \to 0,\ \forall x \in [0,1]$, under known long run variances,
2. proving tightness of $(\check{W}_{N,T}(x) - \check{W}_{N,T}(x))_{x \in [0,1]}$ under known long run variances,
3. verifying $\sup_{x \in [0,1]} |\check{W}_{N,T}(x) - \check{W}_{N,T}(x)| \to 0$ under estimated long run variances.

First denote $E_{N,T}$ the event that $\hat{\sigma}_{i,T} \geq \sigma_i/2$ for $i = 1, \ldots, N$, then

$$
P(E_{N,T}) \geq 1 - \sum_{i=1}^{N} P(|\hat{\sigma}_{i,T} - \sigma_i| > \sigma_i/2) \geq 1 - \frac{N}{T^\alpha} T^\alpha \max_{i \leq N} P(|\hat{\sigma}_{i,T} - \sigma_i| > \sigma_0/2) \to 1
$$

which enables us to assume this case in the following. Let furthermore $F_{N,T}$ denote the event that $|\hat{\mu}_{i,T} - \mu_i| \leq T^{-\beta}$ and $|\frac{1}{\hat{\sigma}_{i,T}} - \frac{1}{\sigma_i}| \leq T^{-\beta}$ for $i = 1, \ldots, N$:

$$
\begin{aligned}
P(F_{N,T}) \geq & 1 - \sum_{i=1}^{N} P(|\hat{\mu}_{i,T} - \mu_i| \geq T^{-\beta}) - \sum_{i=1}^{N} P\left(\left|\frac{\hat{\sigma}_{i,T} - \sigma_i}{\sigma_i \hat{\sigma}_{i,T}}\right| \geq T^{-\beta}\right) \\
\geq & 1 - \frac{N}{T^\alpha} T^\alpha \max_{i \leq N} P(|\hat{\mu}_{i,T} - \mu_i| > T^{-\beta}) - \frac{N}{T^\alpha} T^\alpha \max_{i \leq N} P(|\hat{\sigma}_{i,T} - \sigma_i| > \sigma_0^2/2T^{-\beta}) \to 1
\end{aligned}
$$

and so we can prove Theorem 5 under $F_{N,T}$. Let furthermore w.l.o.g $\mu_i = 0,\ \sigma_i = 1,\ i = 1, \ldots, N$.

Proof of step 1 of Theorem 5. The expansion

$$
\Psi_i\left(\frac{X_{i,t} - \hat{\mu}_{i,T}}{\hat{\sigma}_{i,T}}\right) = \Psi_i\left(X_{i,t} + X_{i,t}\left(\frac{1}{\hat{\sigma}_{i,T}} - 1\right) - \hat{\mu}_{i,T}\left(\frac{1}{\hat{\sigma}_{i,T}} - 1\right) - \hat{\mu}_{i,T}\right)
$$

is almost impossible to work with, since $\hat{\mu}_{i,T}$ and $\hat{\sigma}_{i,T}$ depend on $(X_{i,t})_{t=1,\ldots,T}$. Instead one can look at $\Psi_i\left(X_{i,t} + X_{i,t} d T^{-\beta} + e T^{-\beta}\right) = Z_{i,T}(d,e)$, where d and e are non random with $|d|, |e| \leq 1$, since we are in the case of $F_{N,T}$. Denote

$$
\bar{S}_{d,e}^{(i)}(x) = \frac{1}{\sqrt{T} v_i}\left(\sum_{t=1}^{\lfloor Tx \rfloor} Z_{i,t}(d,e) - \frac{\lfloor Tx \rfloor}{T} \sum_{t=1}^{T} Z_{i,t}(d,e)\right), \quad x \in [0,1]
$$

the disturbed CUSUM-statistic

$$\bar{W}_{N,T,d,e}(x) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(\left(\bar{S}_{d,e}^{(i)}(x) \right)^2 - \frac{\lfloor xT \rfloor (T - \lfloor xT \rfloor)}{T^2} \right), \quad x \in [0,1]$$

the related panel-CUSUM. To prove

$$\sup_{d,e \in [-1,1]} |W_{N,T}(x) - \bar{W}_{N,T,d,e}(x)| \to 0, \quad x \in (0,1)$$

we have to bound the difference between the squared individual CUSUM-statistics

$$\begin{aligned}
&\left(\bar{S}_{d,e}^{(i)}(x) \right)^2 - \left(S^{(i)}(x) \right)^2 \\
&\quad = \frac{1}{\sqrt{T} v_i} \left(\sum_{t=1}^{k} Z_{i,t}(d,e) - \frac{k}{T} \sum_{t=1}^{T} Z_{i,t}(d,e) - \sum_{t=1}^{k} \Psi_i(X_{i,t}) + \frac{k}{T} \sum_{t=1}^{T} \Psi_i(X_{i,t}) \right) \\
&\quad \cdot \frac{1}{\sqrt{T} v_i} \left(\sum_{t=1}^{k} Z_{i,t}(d,e) - \frac{k}{T} \sum_{t=1}^{T} Z_{i,t}(d,e) + \sum_{t=1}^{k} \Psi_i(X_{i,t}) - \frac{k}{T} \sum_{t=1}^{T} \Psi_i(X_{i,t}) \right) = A_i B_i
\end{aligned}$$

where $k = \lfloor Tx \rfloor$. We will show that $\mathbb{E}\left(B_i^2\right)$ is bounded while $\mathbb{E}\left(A_i^2\right)$ converges to 0 sufficiently fast. Essential in both calculations is a Taylor expansion of order m for $Z_{i,t}(d,e)$ around $X_{i,t}$ for $t = 1, \ldots, T$:

$$\begin{aligned}
Z_{i,t}(d,e) = \Psi_i(X_{i,t}) &+ \sum_{r=1}^{m} \frac{\Psi_i(X_{i,t})^{(r)}}{r!} \left(X_{i,t} d T^{-\beta} + e T^{-\beta} \right)^r \\
&+ \frac{\Psi_i(\xi_{i,t})^{(m+1)}}{(m+1)!} \left(X_{i,t} d T^{-\beta} + e T^{-\beta} \right)^{m+1} \qquad (4.32)
\end{aligned}$$

for some $\xi_{i,t} \in [X_{i,t} - X_{i,t}|d|T^{-\beta} - |e|T^{-\beta}, X_{i,t} + X_{i,t}|d|T^{-\beta} + |e|T^{-\beta}]$. Denote

$$U_i^{(r,s)}(k) = \begin{cases} \sum_{t=1}^{k} \Psi_i^{(r)}(X_{i,t}) X_{i,t}^s - \frac{k}{T} \sum_{t=1}^{T} \Psi_i^{(r)}(X_{i,t}) X_{i,t}^s & r = 0, \ldots, m \\ \sum_{t=1}^{k} \Psi_i^{(r)}(\xi_{i,t}) X_{i,t}^s - \frac{k}{T} \sum_{t=1}^{T} \Psi_i^{(r)}(\xi_{i,t}) X_{i,t}^s & r = m+1 \end{cases}, s \le m$$

the non standardized CUSUM statistics which arises from the $r-th$ Taylor-summand in (4.32). In A_i the original CUSUM $U_i^{(0,0)}(k)$ cancels out and by repeated application of the c_r inequality we obtain:

$$\mathbb{E}(A_i^2) \le \sum_{r=1,\ldots,m+1, s \le m} \frac{2^{(m+1)(m+2)/2-1}}{T v_i^2} d^{2s} e^{2(r-s)} T^{-2\beta r} \frac{\binom{r}{s}}{r!} \mathbb{E}\left(\left[U_i^{(r,s)}(k) \right]^2 \right)$$

Because of assumption i) and ii) of Theorem 5 the processes $\left(\Psi_i(X_{i,t})^{(r)} X_{i,t}^s \right)_{t \in \mathbb{N}}$ is also P-NED for $r = 1, \ldots, m$ and $s \le m$, so we can apply the same calculations as in step 1 of the proof of Theorem 3 and receive $\frac{1}{T}\mathbb{E}\left(\left[U_i^{(r,s)}(k) \right]^2 \right) \le C_{63} + \frac{C_64}{T}$ for $r = 1, \ldots, m$. For $m+1$ we use the upper bounds c and d of assumption i) and ii) of Theorem 5 to obtain $\frac{1}{T^2}\mathbb{E}\left(\left[U_i^{(r,s)}(k) \right]^2 \right) \le C_{63}$. So we have $\mathbb{E}(A_i^2) \le C_{65} T^{-2\beta} + 2 C_{64} T^{-2\beta(m+1)+1}$ and by the same calculations $\mathbb{E}(B_i^2) \le C_{66} + C_{67} T^{-2\beta} + C_{68} T^{-2\beta(m+1)+1}$. Finally we apply the Cauchy-Schwarz inequality to obtain

$$\sup_{d,e \in [-1,1]} E(|W_{N,T}(x) - \bar{W}_{N,T,d}(x)|) \le \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(C_{69} T^{-\beta} + C_{70} T^{-(m+1)\beta + 1/2} \right) \to 0$$

which implies pointwise convergence. □

Step 2 of the proof of Theorem 5. Denote $\lfloor Tx\rfloor = u > v = \lfloor Ty\rfloor$, like in step 4 of the proof of Theorem 3, we want to apply the tightness critierion of Billingsley (1968). But before we use the Taylor expansion (4.32)

$$\begin{aligned}
W_{N,T}(x) - \bar{W}_{N,T,d}(x) &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left[\frac{1}{Tv_i^2} U_i^{(0,0)}(u)^2 - \frac{1}{Tv_i^2} \left(\sum_{r=0,\dots,m+1, s\le r} a_{r,s} U_i^{(r,s)}(u) \right)^2 \right] \\
&= \sum_{r=0,\dots,m+1, s\le r} \; \sum_{n=0,\dots,m+1, p\le n} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{1}{Tv_i^2} a_{r,s} a_{n,p} U_i^{(r,s)}(u) U_i^{(n,p)}(u)
\end{aligned}$$

where $a_{r,s} = \frac{d^s e^{r-s} T^{-\beta r} \binom{r}{s}}{r!}$ to see that we only need to prove tightness for the individual summands

$$\left(A_{N,T}(x)^{rsnp}\right)_{x\in[0,1]} = \left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{1}{Tv_i^2} a_{r,s} a_{n,p} U_i^{(r,s)}(\lfloor Tx\rfloor) U_i^{(n,p)}(\lfloor Tx\rfloor) \right)_{x\in[0,1]}$$

for $r, n = 1, \dots, m+1$, $s \le r$ and $p \le n$. Denote

$$M_i^{(rsnp)} = \frac{1}{Tv_i^2} a_{r,s} a_{n,p} U_i^{(r,s)}(u) U_i^{(n,p)}(u) - \frac{1}{Tv_i^2} a_{r,s} a_{n,p} U_i^{(r,s)}(v) U_i^{(n,p)}(u),$$

then we expand the difference of the fourth moment to

$$\begin{aligned}
&\mathbb{E}(|A_{N,T}(x)^{rsnp} - A_{N,T}(y)^{rsnp}|^4) \\
&\quad = \frac{1}{N^2} \sum_{t\ne u\ne v\ne w} \mathbb{E}(M_t^{(rsnp)}) \mathbb{E}(M_u^{(rsnp)}) \mathbb{E}(M_v^{(rsnp)}) \mathbb{E}(M_w^{(rsnp)}) \\
&\quad + \frac{1}{N^2} \sum_{t\ne u\ne v} \mathbb{E}\left(\left[M_t^{(rsnp)} \right]^2 \right) \mathbb{E}(M_u^{(rsnp)}) \mathbb{E}(M_v^{(rsnp)}) + \frac{1}{N^2} \sum_{t\ne u} \mathbb{E}\left(\left[M_t^{(rsnp)} \right]^3 \right) \mathbb{E}(M_u^{(rsnp)}) \\
&\quad + \frac{1}{N^2} \sum_{t\ne u} \mathbb{E}\left(\left[M_t^{(rsnp)} \right]^2 \right) \mathbb{E}\left(\left[M_u^{(rsnp)} \right]^2 \right) + \frac{1}{N^2} \sum_{t=1}^{N} \mathbb{E}\left(\left[M_t^{(rsnp)} \right]^4 \right).
\end{aligned}$$

Exemplarily we look at $\mathbb{E}(M_i^{(jk)})$. If $j, k < m+1$ we can apply Proposition 11 to get

$$\begin{aligned}
|\mathbb{E}(M_i^{(rsnp)})| = \Bigg| \frac{a_{r,s} a_{n,p}}{Tv_i^2} \mathbb{E} \Bigg(&\left[\sum_{t=1}^{\lfloor Tx\rfloor} \Psi_i^{(r)}(X_{i,t}) X_{i,t}^s - \frac{\lfloor Tx\rfloor}{T} \sum_{t=1}^{T} \Psi_i^{(r)}(X_{i,t}) X_{i,t}^s \right] \\
&\cdot \left[\sum_{t=\lfloor Ty\rfloor+1}^{\lfloor Tx\rfloor} \Psi_i^{(n)}(X_{i,t}) X_{i,t}^p - \frac{\lfloor Tx\rfloor - \lfloor Ty\rfloor}{T} \sum_{t=1}^{T} \Psi_i^{(n)}(X_{i,t}) X_{i,t}^p \right] \\
&- \left[\sum_{t=1}^{\lfloor Ty\rfloor} \Psi_i^{(n)}(X_{i,t}) X_{i,t}^p - \frac{\lfloor Ty\rfloor}{T} \sum_{t=1}^{T} \Psi_i^{(n)}(X_{i,t}) X_{i,t}^p \right] \\
&\cdot \left[\sum_{t=\lfloor Ty\rfloor+1}^{\lfloor Tx\rfloor} \Psi_i^{(r)}(X_{i,t}) X_{i,t}^s - \frac{\lfloor Tx\rfloor - \lfloor Ty\rfloor}{T} \sum_{t=1}^{T} \Psi_i^{(r)}(X_{i,t}) X_{i,t}^s \right] \Bigg) \Bigg| \\
&\le \frac{a_{r,s} a_{n,p}}{\delta^2 T} \left(\sqrt{C_{71} \lfloor Tx\rfloor} \sqrt{C_{72}(\lfloor Tx\rfloor - \lfloor Ty\rfloor)} + \sqrt{C_{73} \lfloor Tx\rfloor} \sqrt{C_{74}(\lfloor Tx\rfloor - \lfloor Ty\rfloor)} \right) \\
&\le C_{78} T^{-\beta(r+n)} \sqrt{x-y}
\end{aligned}$$

Analogously one can prove $\mathbb{E}([M_i^{(rsnp)}]^2) \leq T^{-2\beta(n+p)}C_{79}(x-y)$, $|\mathbb{E}([M_i^{rsnp}]^3)| \leq C_{80}T^{-3\beta(j+k)}(x-y)^{\frac{3}{2}}$ and $\mathbb{E}([M_i^{rsnp}]^4) \leq C_{81}T^{-4\beta(n+p)}(x-y)^2$ where non of these bounds depend on d or e, so that

$$\begin{aligned}\sup_{d,e\in[-1,1]} \mathbb{E}(|A_{N,T}(x)^{rsnp} - A_{N,T}(y)^{rsnp}|^4) &\leq C_{82}T^{-4\beta(j+k)}(x-y)^2\left(N^2+N+1+\frac{1}{N}\right)\\ &\leq C_{83}|x-y|^2\end{aligned}$$

which proves tightness. We cannot use Proposition 11 for the Taylor-remainders which occur if $j = m+1$ or $k = m+1$. In this cases we use assumption ii) of Theorem 5 to get

$$\mathbb{E}\left\{\left[\sum_{i=1}^{T}\Psi_i^{(m+1)}(\xi_{i,t})X_{i,t}^s\right]^k\right\} \leq T^k d. \tag{4.33}$$

So if only one of r, n equals $m+1$ we have

$$\begin{aligned}&\sup_{d,e\in[-1,1]} \mathbb{E}(|B_{N,T}(x)^{rs(m+1)p} - B_{N,T}(y)^{rs(m+1)p)}|^4)\\ &\qquad\leq C_{84}T^{-4\beta(r+m+1)+2}(x-y)^2\left(N^2+N+1+\frac{1}{N}\right) \leq C_{85}|x-y|^2\end{aligned}$$

respectively if both are $m+1$

$$\begin{aligned}&\mathbb{E}(|B_{N,T}(x)^{(m+1)s(m+1)p} - B_{N,T}(y)^{(m+1)s(m+1)p}|^4)\\ &\quad\leq C_{86}T^{-8\beta(m+1)+4}(x-y)^2\left(N^2+N+1+\frac{1}{N}\right) \leq C_{87}|x-y|^2.\end{aligned}$$

□

Step 3 of the proof of Theorem 5. We continue to assume that w.l.o.g. $\mu_i = 0$, $\sigma_i = 1$ for $i = 1, \dots, N$. The proof follows that of Theorem 4. We first show that the difference between the long run variance with known standardization $\hat{v}_i$ and the long run variance with estimated standardization

$$\check{v}_i = \check{\gamma}_{i,0} + 2\sum_{i=0}^{b_{i,T}}\check{\gamma}_{i,h}k\left(\frac{h}{b_T}\right),\ i = 1, \dots, N,$$

where

$$\check{\gamma}_i(h) = \frac{1}{T}\sum_{t=1}^{N-h}\left(\Psi_i\left[\frac{X_{i,t}-\hat{\mu}_{i,T}}{\hat{\sigma}_{i,T}}\right] - \check{Y}_i\right)\left(\Psi_i\left[\frac{X_{i,t}-\hat{\mu}_{i,T}}{\hat{\sigma}_{i,T}}\right] - \check{Y}_i\right),\ h = 0, \dots, T-1,$$

and $\check{Y}_i = \frac{1}{T}\sum_{t=1}^{T}\Psi_i\left(\frac{X_{i,t}-\hat{\mu}_{i,T}}{\hat{\sigma}_{i,T}}\right)$ converge with rate $T^{-\min(2\beta, 2\beta(m+1)-1)}$. We use the Taylor series (4.32) and denote $Z_{i,t}^{(k,l)} = \Psi_i(X_{i,t})^{(k)}X_{i,t}^l - \frac{1}{T}\sum_{r=1}^{T}\Psi_i(X_{i,r})^{(k)}X_{i,t}^l$ to obtain

$$\begin{aligned}\hat{v}_i^2 - \check{v}_i^2 &= \frac{1}{T}\sum_{s,t=1}^{T}\left(\Psi_i(X_{i,t}) - \frac{1}{T}\sum_{r=1}^{T}\Psi_i(X_{i,r})\right)\left(\Psi_i(X_{i,s}) - \frac{1}{T}\sum_{r=1}^{T}\Psi_i(X_{i,r})\right)k\left(\frac{s-t}{b_T}\right)\\ &\quad - \sum_{k=0,l\leq k}^{m+1}\ \sum_{n=0,p\leq n}^{m+1} a_{k,l}a_{n,p}\frac{1}{T}\sum_{s.t=1}^{T}Z_{i,t}^{(k,l)}Z_{i,t}^{(n,p)}k\left(\frac{s-t}{b_T}\right).\end{aligned} \tag{4.34}$$

So by (4.34) and the c_r inequality we see

$$\mathbb{E}\left(\left[\hat{v}_i^2 - \check{v}_i^2\right]^2\right) \leq 2^{\frac{(m+2)(m+1)}{2}-1} \sum_{\substack{n,\,k\,=\,0,\,\ldots,\,m+1 \\ n,\,k\,\neq\,(0,0) \\ l\,\leq\,k;\,p\,\leq\,n}} a_{k,l}^2 a_{n,p}^2 \frac{1}{T^2}\mathbb{E}\left(\left[\sum_{s,t=1}^{T} Z_{i,t}^{(k,l)} Z_{i,s}^{(n,p)} k\left(\frac{s-t}{b_T}\right)\right]^2\right).$$

To shorten notation we set $\overline{\Psi(X_{i,t})^{(k)}X_{i,t}^l} = \Psi(X_{i,t})^{(k)}X_{i,t}^l - \mathbb{E}(\Psi(X_{i,1})^{(k)}X_{i,1}^l)$ for $i = 1, \ldots, N$ and $k, l \leq m+1$ and arrive at

$$\begin{aligned}
&\frac{1}{T^2}\mathbb{E}\left(\left[\sum_{s,t=1}^{T} Z_{i,t}^{(k,l)} Z_{i,s}^{(n,p)} k\left(\frac{s-t}{b_T}\right)\right]^2\right) \\
&\quad\leq \frac{2^4}{T^2}\mathbb{E}\left(\left[\sum_{s,t=1}^{T} \overline{\Psi(X_{i,t})^{(k)}X_{i,t}^l}\;\overline{\Psi(X_{i,s})^{(n)}X_{i,s}^p} k\left(\frac{s-t}{b_T}\right)\right]^2\right) \\
&\quad+ \frac{2^4}{T^4}\mathbb{E}\left(\left[\sum_{s,t,r=1}^{T} \overline{\Psi(X_{i,t})^{(k)}X_{i,t}^l}\;\overline{\Psi(X_{i,r})^{(n)}X_{i,r}^p} k\left(\frac{s-t}{b_T}\right)\right]^2\right) \\
&\quad+ \frac{2^4}{T^4}\mathbb{E}\left(\left[\sum_{s,t,r=1}^{T} \overline{\Psi(X_{i,r})^{(k)}X_{i,r}^l}\;\overline{\Psi(X_{i,s})^{(n)}X_{i,s}^p} k\left(\frac{s-t}{b_T}\right)\right]^2\right) \\
&\quad+ \frac{2^4}{T^6}\mathbb{E}\left(\left[\sum_{s,t,r,u=1}^{T} \overline{\Psi(X_{i,r})^{(k)}X_{i,r}^l}\;\overline{\Psi(X_{i,u})^{(n)}X_{i,u}^p} k\left(\frac{s-t}{b_T}\right)\right]^2\right) = A_1 + A_2 + A_3 + A_4.
\end{aligned}$$

We exemplarily look at A_1. Let $r, n < m+1$, like in the proof of Proposition 10 we want to rearrange the summands and split the expectations where the time lag is largest by applying Proposition 9. Here we have the problem that we apply two different functions $g_1(x) = \Psi^{(k)}(x)x^l - \mathbb{E}(\Psi^{(k)}(X_{i,1})X_{i,1}^l)$ and $g_2(x) = \Psi^{(n)}(x)x^p - \mathbb{E}(\Psi^{(n)}(X_{i,1})X_{i,1}^p)$ to the random variables $X_{i,t}$, so after the rearrangement we have six different sums, the one where g_1 is applied to the two random variables with two smallest indices and g_2 to the other and so on:

$$\begin{aligned}
A_1 &\leq 4 \sum_{\substack{a,b,c,d\,\in\,\{1,2\} \\ a+b+c+d\,=\,6}} \sum_{1\leq s\leq t\leq u\leq v\leq T} \left|\mathbb{E}[g_a(X_{i,s}), g_b(X_{i,t}), g_c(X_{i,u}), g_d(X_{i,v})]\right| \\
&\leq 4 \sum_{\substack{a,b,c,d\,\in\,\{1,2\} \\ a+b+c+d\,=\,6}} \sum_{\substack{s,t,u,v\,=\,1 \\ s+t+u+v\,\leq\,T}} \left|\mathbb{E}[g_a(X_{i,s}), g_b(X_{i,s+t}), g_c(X_{i,s+t+u}), g_d(X_{i,s+t+u+v})]\right| \\
&\leq 4T \sum_{\substack{a,b,c,d\,\in\,\{1,2\} \\ a+b+c+d\,=\,6}} \left(\sum_{u,v\leq t} \left|\mathbb{E}[g_a(X_{i,s}), g_b(X_{i,s+t}), g_c(X_{i,s+t+u}), g_d(X_{i,s+t+u+v})]\right|\right. \\
&+ \sum_{t,v\leq u} \left|\mathbb{E}[g_a(X_{i,s}), g_b(X_{i,s+t}), g_c(X_{i,s+t+u}), g_d(X_{i,s+t+u+v})]\right| \\
&+ \left.\sum_{t,u\leq v} \left|\mathbb{E}[g_a(X_{i,s}), g_b(X_{i,s+t}), g_c(X_{i,s+t+u}), g_d(X_{i,s+t+u+v})]\right|\right) = B_1 + B_2 + B_3.
\end{aligned}$$

Proposition 9 yields

$$B_1 \leq C_{88} T \sum_{t=1} t^2 (1+t)^{-b} = TC_{89} \geq B_3$$

and

$$\begin{aligned} B_2 &\leq TC_{90} + C_{91} \sum_{\substack{a,b,c,d \in \{1,2\} \\ a+b+c+d=6}} \sum_{u,v=1}^{T} \mathbb{E}[g_a(X_{i,s}), g_b(X_{i,s+t})] \sum_{u,v=1}^{T} \mathbb{E}[g_c(X_{i,s}), g_d(X_{i,s+t})] \\ &\leq TC_{92} + C_{93} T^2. \end{aligned}$$

If $r = m+1$ or $n = m+1$ one can use (4.33) which yields $A_1 \leq C_{94} T^3$ respectively $A_1 \leq C_{95} T^4$ if $r = n = m+1$. Together we obtain

$$\sup_{d,e \in [-1,1]} \mathbb{E}\left(\left[\hat{v}_i^2 - \check{v}_i^2 \right]^2 \right) \leq C_{96} T^{-2\beta} + C_{97} T^{-2\beta(m+1)+1} + C_{98} T^{-4\beta(m+1)+2},$$

from now on we follow the proof of Theorem 4. □

Proof of Theorem 6. First we show that $I_{\{X_i \leq x\}}$ is P-NED for $x \in \mathbb{R}$ with approximating constants $\tilde{a}_k = a_k^\kappa + \Phi(a_k^\kappa) a_k$ and error function $\tilde{\Phi}(x) = I_{(0,1)}(x)$. Denote therefore $R = \sup_{x \in \mathbb{R}} f(x)$, then

$$\begin{aligned} &P(|I_{\{X_0 \leq x\}} - I_{\{f_k(Z_{-k},\ldots,Z_k) \leq x\}}| > \epsilon) \\ &\qquad = P(|I_{\{X_0 \leq x\}} - I_{\{f_k(Z_{-k},\ldots,Z_k) \leq x\}}| > \epsilon || X_0 - x| \leq a_k^\kappa) P(|X_0 - x| \leq a_k^\kappa) \\ &\qquad + P(|I_{\{X_0 \leq x\}} - I_{\{f_k(Z_{-k},\ldots,Z_k) \geq x\}}| > \epsilon || X_0 - x| \leq a_k^\kappa) P(|X_0 - x| \geq a_k^\kappa) \\ &\qquad \leq 2 a_k^\kappa R + \Phi(a_k^\kappa) a_k. \end{aligned}$$

The above probability is 0 for $\epsilon \geq 1$, so we can choose $\tilde{\Phi}(x) = I_{(0,1)}(x)$ for $x > 0$. Now we follow the proof in Serfling (1980), except for using Markov's inequality in combination with Proposition 11 instead of Hoeffding's inequality. For better readability we abbreviate μ_{med} as μ and σ_{MAD} as σ as well as its empirical versions $\hat{\mu}_{med,T}$ as $\hat{\mu}$ and $\hat{\sigma}_{MAD,T}$ as $\hat{\sigma}$. Let F denote the distribution function of X_1 and $\hat{F}_T$ the empirical distribution of $(X_1, \ldots, X_T)$, then

$$P(|\hat{\mu} - \mu| > \epsilon) = P(\hat{\mu} > \mu + \epsilon) + P(\hat{\mu} < \mu - \epsilon). \tag{4.35}$$

For the first summand in (4.35) one has

$$\begin{aligned} P(\hat{\mu} > \mu + \epsilon) = P(1/2 > \hat{F}_T(\mu + \epsilon)) &= P\left(\sum_{t=1}^{T} I_{\{X_t > \mu + \epsilon\}} > T/2 \right) \\ &\leq P\left(\sum_{t=1}^{T} \left\{ I_{\{X_t > \mu + \epsilon\}} - \mathbb{E}\left[I_{\{X_t > \mu + \epsilon\}} \right] \right\} > T[F(\mu + \epsilon) - 1/2] \right) \\ &\leq \frac{G_2 T^{\lfloor p/2 \rfloor}}{T^p F(\mu + \epsilon - 1/2)^p} \leq \frac{G_2}{T^{p/2} \epsilon^p M^p} \end{aligned}$$

where the last inequality is due to the mean value theorem. The same calculation for the second summand in (4.35) yields (4.8). Now we proof the inequality for the MAD in the same way as the one for the median. Denote G the distribution function of Y_1, then

$$P(|\hat{\sigma} - \sigma| > \epsilon) = P(\hat{\sigma} > \sigma + \epsilon) + P(\hat{\sigma} < \sigma - \epsilon). \tag{4.36}$$

For the first summand in (4.36) we have:

$$
\begin{aligned}
P(\hat{\sigma} > \sigma + \epsilon) &= P\left(\sum_{t=1}^{T} I_{\{Y_i > \sigma+\epsilon\} > T/2}\right) \\
&\leq P\left(\sum_{t=1}^{T} I_{\{|X_t - \hat{\mu}| > \sigma+\epsilon\}} > T/2 \middle| |\hat{\mu} - \mu| < \epsilon/2\right) P(|\hat{\mu} - \mu| < \epsilon/2) \\
&+ P\left(\sum_{t=1}^{T} I_{\{|X_t - \hat{\mu}| > \sigma+\epsilon\}} > T/2 \middle| |\hat{\mu} - \mu| > \epsilon/2\right) P(|\hat{\mu} - \mu| > \epsilon/2) \\
&\leq P\left(\sum_{t=1}^{T} I_{\{|X_t - \hat{\mu}| > \sigma+\epsilon/2\}} > T/2\right) + P(|\hat{\mu} - \mu| > \epsilon/2) \\
&\leq \frac{\tilde{G}_2}{T^{p/2}(\epsilon/2)^p \tilde{M}^p} + \frac{G_2}{T^{p/2}(\epsilon/2)^p M^p}
\end{aligned}
$$

which one also obtains for the second summand in (4.36). □

Bibliography

Al-Homidan, S., 2006. Semidefinite and second-order cone optimization approach for the Toeplitz matrix approximation problem. Journal of Numerical Mathematics 14, 1–15.

Anderson, T., 2003. An introduction to multivariate statistical analysis. J. Wiley, Hoboken. 3rd edition.

Anderson, T.W., 1971. The statistical analysis of time series. John Wiley & Sons.

Andrews, D.W., 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica: Journal of the Econometric Society , 817–858.

Aston, J.A.D., Kirch, C., 2014. Efficiency of change point tests in high dimensional settings. ArXiv e-prints `1409.1771`.

Aue, A., Hörmann, S., Horváth, L., Reimherr, M., 2009. Break detection in the covariance structure of multivariate time series models. The Annals of Statistics 37, 4046–4087.

Bahadur, R.R., 1966. A note on quantiles in large samples. The Annals of Mathematical Statistics 37, 577–580.

Bai, J., 2010. Common breaks in means and variances for panel data. Journal of Econometrics 157, 78–92.

Bai, J., Carrion-I-Silvestre, J.L., 2009. Structural changes, common stochastic trends, and unit roots in panel data. The Review of Economic Studies 76, 471–501.

Baillie, R.T., Chung, H., 2001. Estimation of GARCH models from the autocorrelations of the squares of a process. Journal of Time Series Analysis 22, 631–650.

Bali, J.L., Boente, G., Tyler, D.E., Wang, J.L., 2011. Robust functional principal components: A projection-pursuit approach. The Annals of Statistics 39, 2852–2882.

Baltagi, B.H., Feng, Q., Kao, C., 2016. Estimation of heterogeneous panels with structural breaks. Journal of Econometrics 191, 176–195.

Baltagi, B.H., Kao, C., Liu, L., 2017. Estimation and identification of change points in panel models with nonstationary or stationary regressors and error term. Econometric Reviews 36, 85–102.

Basrak, B., Davis, R.A., Mikosch, T., 2002. Regular variation of GARCH processes. Stochastic Processes and their Applications 99, 95–115.

Bernoulli, D., 1777. Dijudicatio maxime probabilis plurium observationum discrepantium atque versimillima inductio inde formanda. Acta Academiae Petropolitanae 1, 3–33. Englisch translation by C. G. Allen in Biometrica 48, 3 –13.

Bickel, P., Doksum, K., 2001. Mathematical statistics. Prentice Hall.

Bierens, H.J., 1981. Robust methods and asymptotic theory in nonlinear econometrics. Springer.

Billingsley, P., 1968. Convergence of probability measures. John Wiley & Sons.

Bilodeau, M., Brenner, D., 1999. Theory of Multivariate Statistics. Springer Texts in Statistics, Springer.

Blomqvist, N., 1950. On a measure of dependence between two random variables. The Annals of Mathematical Statistics 21, 593–600.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31, 307–327.

Bollerslev, T., Chou, R.Y., Kroner, K.F., 1992. ARCH modeling in finance: A review of the theory and empirical evidence. Journal of Econometrics 52, 5–59.

Borovkova, S., Burton, R., Dehling, H., 2001. Limit theorems for functionals of mixing processes with applications to U-statistics and dimension estimation. Transactions of the American Mathematical Society 353, 4261–4318.

Boudt, K., Cornelissen, J., Croux, C., 2012. The Gaussian rank correlation estimator: robustness properties. Statistics and Computing 22, 471–483.

Box, G.E., 1953. Non-normality and tests on variances. Biometrika 40, 318–335.

Box, G.E.P., Jenkins, G.M., Reinsel, G.C., 1994. Time series analysis: forecasting and control. Prentice Hall. 3rd edition.

Bramati, M.C., Croux, C., 2007. Robust estimators for the fixed effects panel data model. The Econometrics Journal 10, 521–540.

Brockwell, P., Davis, R.A., 2006. Time series: theory and methods. Springer, New York. 2nd edition.

Brockwell, P.J., 2009. Autocovariance. Wiley Interdisciplinary Reviews: Computational Statistics 1, 187–198.

Brockwell, P.J., 2011. Autoregressive processes. Wiley Interdisciplinary Reviews: Computational Statistics 3, 316–331.

Bustos, O.H., Yohai, V.J., 1986. Robust estimates for ARMA models. Journal of the American Statistical Association 81, 155–168.

Caiado, J., Crato, N., Peña, D., 2006. A periodogram-based metric for time series classification. Computational Statistics & Data Analysis 50, 2668–2684.

Chakhchoukh, Y., 2010. A new robust estimation method for ARMA models. IEEE Transactions on Signal Processing 58, 3512–3522.

Chan, W.S., 1992. A note on time series model specification in the presence of outliers. Journal of Applied Statistics 19, 117–124.

Chan, W.S., Wei, W.W., 1992. A comparison of some estimators of time series autocorrelations. Computational Statistics & Data Analysis 14, 149–163.

Chang, C.C., Politis, D.N., 2014. Robust autocorrelation estimation. Journal of Computational and Graphical Statistics .

Crainiceanu, C.M., Vogelsang, T.J., 2007. Nonmonotonic power for tests of a mean shift in a time series. Journal of Statistical Computation and Simulation 77, 457–476.

Croux, C., Dehon, C., 2010. Influence functions of the Spearman and Kendall correlation measures. Statistical Methods & Applications 19, 497–515.

Croux, C., Dehon, C., Yadine, A., 2010. The k-step spatial sign covariance matrix. Advances in Data Analysis and Classification 4, 137–150.

Croux, C., Haesbroeck, G., 1999. Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. Journal of Multivariate Analysis 71, 161–190.

Croux, C., Ollila, E., Oja, H., 2002. Sign and rank covariance matrices: statistical properties and application to principal components analysis, in: Dodge, Y. (Ed.), Statistical Data Analysis Based on the L_1-Norm and Related Methods (Papers of the 4th international conference on statistical analysis on the L_1-norm and related methods, Neuchâtel, Switzerland, August 4–9, 2002). Birkhäuser, Basel, pp. 257–269.

Csörgo, M., Horváth, L., 1987. Nonparametric tests for the changepoint problem. Journal of Statistical Planning and Inference 17, 1–9.

Davies, P., 1987. Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. The Annals of Statistics 15, 1269–1292.

Davis, R., Resnick, S., 1986. Limit theory for the sample covariance and correlation functions of moving averages. The Annals of Statistics 14, 533–558.

Davis, R.A., Mikosch, T., 1998. The sample autocorrelations of heavy-tailed processes with applications to ARCH. The Annals of Statistics 26, 2049–2080.

Davydov, Y.A., 1970. The invariance principle for stationary processes. Theory of Probability & Its Applications 15, 487–498.

De Wachter, S., Tzavalis, E., 2005. Monte Carlo comparison of model and moment selection and classical inference approaches to break detection in panel data models. Economics Letters 88, 91–96.

De Wachter, S., Tzavalis, E., 2012. Detection of structural breaks in linear dynamic panel data models. Computational Statistics & Data Analysis 56, 3020–3034.

Dehling, H., Fried, R., Garcia, I., Wendler, M., 2015. Change-point detection under dependence based on two-sample U-statistics, in: Asymptotic Laws and Methods in Stochastics. Springer, pp. 195–220.

Dehling, H., Fried, R., Wendler, M., 2015. A Robust Method for Shift Detection in Time Series. ArXiv e-prints 1506.03345.

Dehling, H., Vogel, D., Wendler, M., Wied, D., 2017. Testing for changes in Kendall's tau. To appear in: Econometric Theory.

Denby, L., Martin, R.D., 1979. Robust estimation of the first-order autoregressive parameter. Journal of the American Statistical Association 74, 140–146.

Deo, C.M., 1973. A note on empirical processes of strong-mixing sequences. The Annals of Probability , 870–875.

Deutsch, S.J., Richards, J.E., Swain, J.J., 1990. Effects of a single outlier on ARMA identification. Communications in Statistics: Theory and Methods 19, 2207–2227.

Donoho, D.L., 1982. Breakdown properties of multivariate location estimators. Ph.D. thesis. Harvard University. http://statweb.stanford.edu/%7Edonoho/Reports/Oldies/BPMLE.pdf.

Duan, J.C., 1995. The GARCH option pricing model. Mathematical Finance 5, 13–32.

Dürre, A., Fried, R., 2016. Robust change-point detection in panel data. ArXiv e-prints 1611.02571.

Dürre, A., Fried, R., Liboschik, T., 2015a. Robust estimation of (partial) autocorrelation. Wiley Interdisciplinary Reviews: Computational Statistics 7, 205–222.

Dürre, A., Fried, R., Liboschik, T., Rathjens, J., 2016. robts: Robust Time Series Analysis. R package version 0.2.2.

Dürre, A., Fried, R., Vogel, D., 2017. The spatial sign covariance matrix and its application for robust correlation estimation. Austrian Journal of Statistics 46, 13–22.

Dürre, A., Tyler, D.E., Vogel, D., 2016. On the eigenvalues of the spatial sign covariance matrix in more than two dimensions. Statistics & Probability Letters 111, 80–85.

Dürre, A., Vogel, D., 2016a. Asymptotics of the two-stage spatial sign correlation. Journal of Multivariate Analysis 144, 54–67.

Dürre, A., Vogel, D., 2016b. sscor: Spatial sign correlation. R package version 0.2.

Dürre, A., Vogel, D., Fried, R., 2015b. Spatial sign correlation. Journal of Multivariate Analysis 135, 89–105.

Dürre, A., Vogel, D., Tyler, D.E., 2014. The spatial sign covariance matrix with unknown location. Journal of Multivariate Analysis 130, 107–117.

Eden, T., Yates, F., 1933. On the validity of Fisher's z test when applied to an actual example of non-normal data. The Journal of Agricultural Science 23, 6–17.

Ferretti, N.E., Kelmansky, D.M., Yohai, V.J., 1991. Estimators based on ranks for ARMA models. Communications in Statistics – Theory and Methods 20, 3879–3907.

Filzmoser, P., Fritz, H., Kalcher, K., 2011. pcaPP: Robust PCA by Projection Pursuit. R package version 1.9-44.

Fisher, R.A., 1921. On the probable error of a coefficient of correlation deduced from a small sample. Metron 1, 3–32.

Fox, A.J., 1972. Outliers in time series. Journal of the Royal Statistical Society. Series B (Methodological) 34, 350–363.

Frahm, G., 2004. Generalized elliptical distributions: theory and applications. Ph.D. thesis. Universität zu Köln.

Fraiman, R., Muniz, G., 2001. Trimmed means for functional data. Test 10, 419–440.

Fried, R., Liboschik, T., Elsaied, H., Kitromilidou, S., Fokianos, K., 2014. On outliers and interventions in count time series following GLMs. Austrian Journal of Statistics 43, 181–193.

Garel, B., Hallin, M., 1999. Rank-based autoregressive order identification. Journal of the American Statistical Association 94, 1357–1371.

Gather, U., Bauer, M., Fried, R., 2002. The identification of multiple outliers in online monitoring data. Estadistica 54, 289–338.

Genton, M.G., Ma, Y., 1999. Robustness properties of dispersion estimators. Statistics & Probability Letters 44, 343–350.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Hothorn, T., 2014. mvtnorm: Multivariate Normal and t Distributions. R package version 0.9-9997.

Gervini, D., 2003. A robust and efficient adaptive reweighted estimator of multivariate location and scatter. Journal of Multivariate Analysis 84, 116–144.

Gervini, D., 2008. Robust functional estimation using the median and spherical principal components. Biometrika 95, 587–600.

Ghosh, J.K., 1971. A new proof of the Bahadur representation of quantiles and an application. The Annals of Mathematical Statistics 42, 1957–1961.

Gnanadesikan, R., Kettenring, J.R., 1972. Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics 28, 81–124.

Gradshteyn, I., Ryzhik, I., 2000. Table of integrals, series, and products. Translation edited and with a preface by Alan Jeffrey and Daniel Zwillinger. Amsterdam: Elsevier/Academic Press. 6th edition.

Hadi, A.S., Imon, A., Werner, M., 2009. Detection of outliers. Wiley Interdisciplinary Reviews: Computational Statistics 1, 57–70.

Haldane, J., 1948. Note on the median of a multivariate distribution. Biometrika 35, 414–417.

Hampel, F.R., 1968. Contribution to the theory of robust estimation. Ph. D. Thesis, University of California, Berkeley .

Hampel, F.R., 1971. A general qualitative definition of robustness. The Annals of Mathematical Statistics , 1887–1896.

Hampel, F.R., 1975. Beyond location parameters: Robust concepts and methods. Bulletin of the International Statistical Institute 46, 375–382.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A., 1986. Robust statistics. The approach based on influence functions. Wiley.

Hey, G., 1938. A new method of experimental sampling illustrated on certain non-normal populations. Biometrika 30, 68–80.

Hodges Jr, J.L., 1967. Efficiency in normal samples and tolerance of extreme values for some estimates of location, in: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, pp. 163–186.

Hoeffding, W., 1948. A class of statistics with asymptotically normal distribution. The Annals of Mathematical Statistics 19, 293–325.

Hörmann, S., 2008. Augmented GARCH sequences: Dependence structure and asymptotics. Bernoulli 14, 543–561.

Horváth, L., Hušková, M., 2012. Change-point detection in panel data. Journal of Time Series Analysis 33, 631–648.

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

Huber, P.J., 1973. Robust regression: asymptotics, conjectures and monte carlo. The Annals of Statistics , 799–821.

Huber, P.J., Ronchetti, E.M., 2009. Robust statistics. Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley. 2nd edition.

Hubert, M., Debruyne, M., 2009. Breakdown value. Wiley Interdisciplinary Reviews: Computational Statistics 1, 296–302.

Hubert, M., Debruyne, M., 2010. Minimum covariance determinant. Wiley Interdisciplinary Reviews: Computational Statistics 2, 36–43.

Hušková, M., 1996. Tests and estimators for the change point problem based on M-statistics. Statistics & Risk Modeling 14, 115–136.

Hušková, M., Marušiaková, M., 2012. M-procedures for detection of changes for dependent observations. Communications in Statistics-Simulation and Computation 41, 1032–1050.

Ibragimov, I.A., 1962. Some limit theorems for stationary processes. Theory of Probability & Its Applications 7, 349–382.

Im, K.S., Lee, J., Tieslau, M., 2005. Panel lm unit-root tests with level shifts. Oxford Bulletin of Economics and Statistics 67, 393–419.

Jirak, M., et al., 2015. Uniform change point tests in high dimension. The Annals of Statistics 43, 2451–2483.

Joseph, L., Wolfson, D.B., 1992. Estimation in multi-path change-point problems. Communications in Statistics-Theory and Methods 21, 897–913.

Kallabis, R.S., Neumann, M.H., et al., 2006. An exponential inequality under weak dependence. Bernoulli 12, 333–350.

Kano, Y., 1994. Consistency property of elliptic probability density functions. Journal of Multivariate Analysis 51, 139–147.

Karavias, Y., Tzavalis, E., 2012. Generalized fixed-T panel unit root tests allowing for structural breaks. Discussion paper. Granger Centre Discussion Paper No. 12/02, University of Nottingham.

Kemperman, J.H.B., 1987. The median of a finite measure on a Banach space, in: Dodge, Y. (Ed.), Statistical Data Analysis Based on the L_1-Norm and Related Methods, Amsterdam: North-Holland. pp. 217–230.

Kent, J.T., Tyler, D.E., 1996. Constrained M-estimation for multivariate location and scatter. The Annals of Statistics 24, 1346–1370.

Kiefer, J., 1967. On Bahadur's representation of sample quantiles. The Annals of Mathematical Statistics 38, 1323–1342.

Kim, D., 2011. Estimating a common deterministic time trend break in large panels with cross sectional dependence. Journal of Econometrics 164, 310–330.

Koenker, R., 2005. Quantile regression. Cambridge university press.

Koltchinskii, V., Dudley, R., 2000. On spatial quantiles., in: Korolyuk, V. et al. (Ed.), Skorokhod's ideas in probability theory., Kiev: Institute of Mathematics of NAS of Ukraine. Proc. Inst. Math. Natl. Acad. Sci. Ukr., Math. Appl. 32. pp. 195–210.

Kontorovich, L.A., Ramanan, K., 2008. Concentration inequalities for dependent random variables via the martingale method. The Annals of Probability 36, 2126–2158.

Koziol, J.A., 1978. Multivariate signed rank statistics for shift alternatives. Statistics: A Journal of Theoretical and Applied Statistics 9, 549–562.

Lamoureux, C.G., Lastrapes, W.D., 1990. Heteroskedasticity in stock return data: volume versus GARCH effects. The Journal of Finance 45, 221–229.

Locantore, N., Marron, J., Simpson, D., Tripoli, N., Zhang, J., Cohen, K., 1999. Robust principal component analysis for functional data. Test 8, 1–73.

Loève, M., 1977. Probability Theory I. Springer-Verlag, New York. 4th edition.

Lopuhaä, H.P., 1989. On the relation between S-estimators and M-estimators of multivariate location and covariance. The Annals of Statistics 17, 1662–1683.

Loretan, M., Phillips, P.C., 1994. Testing the covariance stationarity of heavy-tailed time series: An overview of the theory with applications to several financial datasets. Journal of Empirical Finance 1, 211–248.

Ma, Y., Genton, M.G., 2000. Highly robust estimation of the autocovariance function. Journal of Time Series Analysis 21, 663–684.

Ma, Y., Genton, M.G., 2001. Highly robust estimation of dispersion matrices. Journal of Multivariate Analysis 78, 11–36.

Magnus, J.R., Neudecker, H., 1999. Matrix differential calculus with applications in statistics and econometrics. Wiley Series in Probability and Statistics. Chichester: Wiley. 2nd edition.

Magyar, A., Tyler, D.E., 2011. The asymptotic efficiency of the spatial median for elliptically symmetric distributions. Sankhya B 73, 165–192.

Magyar, A.F., Tyler, D.E., 2014. The asymptotic inadmissibility of the spatial sign covariance matrix for elliptically symmetric distributions. Biometrika 101, 673–688.

Makhoul, J., 1981. Lattice methods in spectral estimation. Applied Time Series Analysis 2, 301–326.

Mallows, C.L., 1975. On some topics in robustness. Unpublished memorandum, Bell Telephone Laboratories, Murray Hill, NJ .

Marden, J.I., 1999. Some robust estimates of principal components. Statistics & Probability Letters 43, 349–359.

Maronna, R.A., 1976. Robust M-estimators of multivariate location and scatter. The Annals of Statistics 4, 51–67.

Maronna, R.A., Martin, R.D., Yohai, V.J., 2006. Robust statistics. J. Wiley, Chichester.

Maronna, R.A., Yohai, V.J., 1976. Robust estimation of multivariate location and scatter. Wiley StatsRef: Statistics Reference Online .

Maronna, R.A., Zamar, R.H., 2002. Robust estimates of location and dispersion for high-dimensional datasets. Technometrics 44.

Martin, R.D., Thomson, D.J., 1982. Robust-resistant spectrum estimation. Proceedings of the IEEE 70, 1097–1115.

Martin, R.D., Yohai, V.J., 1986. Influence functionals for time series. The Annals of Statistics , 781–818.

Masarotto, G., 1987. Robust identification of autoregressive moving average models. Applied statistics 36, 214–220.

Masreliez, C., 1975. Approximate non-gaussian filtering with linear state and observation relations. IEEE Transactions on Automatic Control 20, 107–110.

Maxima, 2014. Maxima, a Computer Algebra System. Version 5.34.1. http://maxima.sourceforge.net/.

Milasevic, P., Ducharme, G., 1987. Uniqueness of the spatial median. The Annals of Statistics 15, 1332–1333.

Mosteller, F., 1946. On some useful 'inefficient' statistics. The Annals of Mathematical Statistics 17, 377–408.

Möttönen, J., Koivunen, V., Oja, H., 1999. Robust autocovariance estimation based on sign and rank correlation coefficients, in: Proceedings of the IEEE Signal Processing Workshop on Higher-Order Statistics, IEEE. pp. 187–190.

Möttönen, J., Oja, H., Tienari, J., 1997. On the efficiency of multivariate spatial sign and rank tests. The Annals of Statistics 25, 542–552.

Nevalainen, J., Larocque, D., Oja, H., 2007. On the multivariate spatial median for clustered data. Canadian Journal of Statistics 35, 215–231.

Nordhausen, K., Oja, H., 2011. Multivariate L_1 methods: The package MNM. Journal of Statistical Software 43, 1–28.

Nordhausen, K., Sirkia, S., Oja, H., Tyler, D.E., 2012. ICSNP: Tools for Multivariate Nonparametrics. R package version 1.0-9.

Oja, H., 2010. Multivariate nonparametric methods with R. An approach based on spatial signs and ranks. Lecture Notes in Statistics 199, New York: Springer.

Paindaveine, D., 2008. A canonical definition of shape. Statistics & Probability Letters 78, 2240–2247.

Paindaveine, D., Verdebout, T., et al., 2016. On high-dimensional sign tests. Bernoulli 22, 1745–1769.

Parzen, E., 1957. On consistent estimates of the spectrum of a stationary time series. The Annals of Mathematical Statistics , 329–348.

Pearson, E.S., 1931. The analysis of variance in cases of non-normal variation. Biometrika 23, 114–133.

Pearson, K., 1907. Mathematical contributions to the theory of evolution. XVI. On further methods of determining correlation. Drapers' company research memoirs: Biometric series, London: Dulau & Co.

Philipp, W., 1986. Invariance principles for independent and weakly dependent random variables, in: Dependence in Probability and Statistics (Proc. Conf. Oberwolfach. 1985). Boston, pp. 225–268.

Politis, D.N., 2009. Financial time series. Wiley Interdisciplinary Reviews: Computational Statistics 1, 157–166.

Politis, D.N., 2011. Higher-order accurate, positive semidefinite estimation of large-sample covariance and spectral density matrices. Econometric Theory 27, 703–744.

Prucha, I., Pötscher, B., 1997. Dynamic nonlinear econometric models: asymptotic theory. Springer.

Quessy, J.F., Saïd, M., Favre, A.C., 2013. Multivariate Kendall's tau for change-point detection in copulas. Canadian Journal of Statistics 41, 65–82.

R Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.

Ramsey, F.L., 1974. Characterization of the partial autocorrelation function. The Annals of Statistics 2, 1296–1301.

Rojo, J., 2013. Heavy-tailed densities. Wiley Interdisciplinary Reviews: Computational Statistics 5, 30–40.

Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., Maechler, M., 2014. robustbase: Basic Robust Statistics. R package version 0.90-2.

Rousseeuw, P.J., 1985. Multivariate estimation with high breakdown point., in: Grossmann, W., Pflug, G.C., Vincze, I., Wertz, W. (Eds.), Mathematical statistics and applications, Proc. 4th Pannonian Symp. Math. Stat., Bad Tatzmannsdorf, Austria, September 4-10, 1983, Vol. B, Dordrecht etc.: D. Reidel. pp. 283–297.

Rousseeuw, P.J., Croux, C., 1993. Alternatives to the median absolute deviation. Journal of the American Statistical Association 88, 1273–1283.

Rousseeuw, P.J., Driessen, K.V., 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212–223.

Salibian-Barrera, M., Yohai, V.J., 2006. A fast algorithm for S-regression estimates. Journal of Computational and Graphical Statistics 15.

Schlittgen, R., Streitberg, B.H.J., 2001. Zeitreihenanalyse. Oldenbourg, München.

Sen, P.K., 1968. Asymptotic normality of sample quantiles for m-dependent processes. The Annals of Mathematical Statistics , 1724–1730.

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

Serfling, R.J., 1980. Approximation theorems of mathematical statistics. John Wiley & Sons.

Serfling, R.J., 1984. Generalized L-, M-, and R-statistics. The Annals of Statistics , 76–86.

Siegel, A.F., 1982. Robust regression using repeated medians. Biometrika 69, 242–244.

Sirkiä, S., Taskinen, S., Oja, H., Tyler, D.E., 2009. Tests and estimates of shape based on spatial signs and ranks. Journal of Nonparametric Statistics 21, 155–176.

Stahel, W., 1981. Robust estimation: Infinitesimal optimality and covariance matrix estimators. Ph.D. thesis. ETH Zürich.

Stigler, S.M., 1973. Simon Newcomb, Percy Daniell, and the history of robust estimation 1885–1920. Journal of the American Statistical Association 68, 872–879.

Taskinen, S., Croux, C., Kankainen, A., Ollila, E., Oja, H., 2006. Influence functions and efficiencies of the canonical correlation and vector estimates based on scatter and shape matrices. Journal of Multivariate Analysis 97, 359–384.

Taskinen, S., Kankainen, A., Oja, H., 2003. Sign test of independence between two random vectors. Statistics & Probability Letters 62, 9–21.

Todorov, V., Filzmoser, P., 2009. An object-oriented framework for robust multivariate analysis. Journal of Statistical Software 32, 1–47.

Tukey, J.W., 1960. A survey of sampling from contaminated distributions. Contributions to Probability and Statistics 2, 448–485.

Tyler, D., 2002. High breakdown point multivariate M-estimation. Estadística 54, 213–247.

Tyler, D.E., 1987. A distribution-free M-estimator of multivariate scatter. The Annals of Statistics 15, 234–251.

Vakili, K., Schmitt, E., 2014. Finding multivariate outliers with fastPCS. Computational Statistics & Data Analysis 69, 54–66.

Vardi, Y., Zhang, C.H., 2001. A modified Weiszfeld algorithm for the Fermat-Weber location problem. Mathematical Programming 90, 559–566.

Vecchia, A., Ballerini, R., 1991. Testing for periodic autocorrelations in seasonal time series data. Biometrika 78, 53–63.

Visuri, S., Koivunen, V., Oja, H., 2000. Sign and rank covariance matrices. Journal of Statistical Planning and Inference 91, 557–575.

Visuri, S., Oja, H., Koivunen, V., 2001. Subspace-based direction-of-arrival estimation using nonparametric statistics. IEEE Transactions on Signal Processing 49, 2060–2073.

Vogel, D., Fried, R., 2011. Elliptical graphical modelling. Biometrika 98, 935–951.

Vogel, D., Fried, R., 2015. Robust change detection in the dependence structure of multivariate time series, in: Modern Nonparametric, Robust and Multivariate Methods. Springer, pp. 265–288.

Vogel, D., Köllmann, C., Fried, R., 2008. Partial correlation estimates based on signs, in: Proceedings of the 1st Workshop on Information Theoretic Methods in Science and Engineering. TICSP series, pp. 1–8.

Vogelsang, T.J., 1999. Sources of nonmonotonic power when testing for a shift in mean of a dynamic time series. Journal of Econometrics 88, 283–299.

Wei, W.W., 1990. Time series analysis: Univariate and Multivariate Methods. Addison-Wesley, Redwood City.

Wendler, M., 2011. Bahadur representation for U-quantiles of dependent data. Journal of Multivariate Analysis 102, 1064–1079.