FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

DOCTORAL THESIS

Přemysl Bejda

Median in some statistical methods

Department of Probability and Mathematical Statistics

Supervisor of the doctoral thesis:	prof. RNDr. Tomáš Cipra, DrSc.
Study programme:	Mathematics
Specialization:	Probability and Statistics, Econometrics and Financial Mathematics

Prague 2017

I would like to thank my supervisor prof. Tomáš Cipra for the guidance of this dissertation, to Lukáš Adam for the help with the second part of the work and to my family for patience and tolerance. I would like to also mention the people who encouraged me to finish the work namely Martina Kadlecová and Michal Šindler.

I declare that I carried out this doctoral thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

In .... date .... signature of the author

Název práce: Median pro různé statistické metody

Autor: Mgr. Přemysl Bejda

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí disertační práce: prof. RNDr. Tomáš Cipra, DrSc., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: V této práci se zaměřujeme na využití robustních vlastností mediánu. Pro algoritmy, které jsou v práci navržené, zkoumáme jejich breakdown point, ale i další vlastnosti jako konsistenci (silnou nebo slabou), ekvivarianci a výpočetní složitost. Z praktických důvodů hledáme především metody, které se snaží najít rovnováhu mezi výpočetní složitostí a dobrými robustními vlastnostmi, protože tyto vlastnosti obvykle stojí proti sobě. Disertace je rozdělena do dvou částí.

V první části navrhujeme robustní metody na bázi exponenciálního vyrovnávání. Nejprve zobecňujeme dřívější výsledky pro exponenciální vyrovnávání v absolutní normě s využitím. regresních kvantyů. Dále navrhujeme metodu založenou na znaménkovém testu, která se snaží vypořádat nejen s odlehlými pozorováními, ale i detekovat čas změny modelu.

V druhé části navrhujeme nové odhady parametru polohy. Konstruujeme je tak, ze nejprve najdeme množinu robustních bodů okolo geometrického mediánu, tuto množiinu dále rozšiřujeme a z bodů této množiny počítáme iterativně vážený průměrr. Díky tomu získáme robustní odhad ve smyslu breakdown pointu, který využívá více informace z pozorovaných hodnot než běžné robustní odhady. Tento přístup se uplatní při konstrukci boxplotu a bagplotu. Odhady konstruujeme na obecném normovaném vektorovém prostoru s tím, že díky využití multifunkcí mohou být tyto odhady definovány jako množiny.

Klíčová slova: Break down point, Parametr polohy, Rekurzivně adaptivní metody, Robustní statistika

Title: Median in some statistical methods

Author: Mgr. Přemysl Bejda

Department: Department of Probability and Mathematical Statistics

Supervisor: prof. RNDr. Tomáš Cipra, DrSc., Department of Probability and Mathematical Statistics

Abstract: This work is focused on utilization of robust properties of median. We propose variety of algorithms with respect to their breakdown point. In addition, other properties are studied such as consistency (strong or weak), equivariance and computational complexity. From practical point of view we are looking for methods balancing good robust properties and computational complexity, because these two properties do not usually correspond to each other. The dissertation is divided to two parts.

In the first part, robust methods similar to the exponential smoothing are suggested. Firstly, the previous results for the exponential smoothing with absolute norm are generalized using the regression quantiles. Further, the method based on the classical sign test is introduced, which deals not only with outliers but also detects change points.

In the second part we propose new estimators of location. These estimators select a robust set around the geometric median, enlarge it and compute the (iterative) weighted mean from it. In this way we obtain a robust estimator in the sense of the breakdown point which exploits more information from observations than standard estimators. We apply our approach on the concepts of boxplot and bagplot. We work in a general normed vector space allowing multi-valued estimators.

Keywords: Break down point, Location parameter, Recursive adaptive methods, Robust statistic

Contents

1. Introduction

Classical statistical methods were developed in the period when the most crucial property of the method was its simplicity. This was mainly because of the lack of modern computational tools. Without them the usage of more complicated techniques could tend toward a higher probability of errors. There are also other advantages of the classical methods over the more complicated ones (or in our case robust) such as easier understanding, more developed theoretical framework or more general applicability (simpler methods are applicable in more situations than highly specialized ones). On the other side these methods are the most suitable only under very restrictive assumptions. Robust statistical methods try to weaken them. E.g., in many statistical models it is assumed that residuals are independent and identically distributed. Moreover, the assumption of normality of residuals is often added and the Euclidean norm is usually employed. The normality of observations can be also justified by theoretical arguments, namely central limit theorem. But for real data these assumptions are often violated.

This naturally leads to the robust statistic which concerns cases when the assumptions of the classical methods are not fulfilled. Its development is enabled by progress in computer science, as the robust methods are usually much more complicated and computationally demanding than the classical methods. As a standard example of the robust technique can serve replacing the l_2 norm by l_1 norm or the removal of observations which does not seem to follow the general pattern (e.g., the observations too far from the rest of the observations etc.).

Literature dealing with the robust statistic is already quite vast. We mention several monographs concerning the field (Maronna et al., 2006), (Jurečková and Picek, 2005) or (Huber and Ronchetti, 2009), where we can find many robust methods suitable in different situations.

From the comparison of different methods it is only clear that none of them is superior to another and that we have to understand the differences between them to utilize the most appropriate one for the problem under investigation. The understanding of the methods is also essential in frequent cases, when they give different results.

We demonstrate now the difference between classical and robust methods on a simple example. Let X stand for the matrix of known regressors, y stand for a vector of regressands, ε stand for a vector of unobserved random variables which elements are i.i.d. (independent identically distributed), normally distributed, with expected value 0, variance σ^2 (i.e., $\varepsilon_t \sim \mathrm{N}(0, \sigma^2)$) and β is a vector of parameters. We consider the model in the form

$$y = X\beta + \varepsilon. \tag{1.1}$$

According to Gauss-Markov theorem, the best linear unbiased estimator is the ordinary least square estimator which has the form $(X^\top X)^{-1} X^\top y$. Here "the best" stands for the fact that the ordinary least square estimator has the lowest variance among the all unbiased linear estimators. To derive that we do not need all the above mentioned assumptions. Namely we do not need to utilize normality and instead of independent and identically distributed residuals it suffices for matrix ε_t to be uncorrelated with each other, to have an expected value equal to zero, to be homoscedastic with finite variance and for X to have a full rank, see (Seber and Lee, 2003). The simple mean is also an example of the least square estimator.

In real data problems it quite often appears that an assumed normality describes the majority of observations; nevertheless, some of the observations can violate this assumption and behave in a different way to the rest. They can follow a different pattern, no pattern at all, they can have infinite variance etc. These observations can be sometimes identified such; that they lie outside the cloud of the rest of the data. Such a behavior is typical across the broad spectrum of datasets. The observations which do not fulfill the assumptions ("behavior") of the rest of the data are called *outliers*. The outliers may have a large distorting effect on classical statistical estimators. The value of the classical estimate influenced by outliers can be misleading.

A different view on data considers its distribution function. We suppose continuous distribution of observations in a random sample and denote $f(x)$ their density function. For the sake of simplicity we further consider $f(x)$ to be symmetric. If for $|x| \to \infty$ the $f(x)$ converges to 0 more slowly than normal density function, we say that the distribution of the random sample is *heavy tailed*. Also in this case the classical models based on the least squares are usually inappropriate.

We return to our example. We have shown that mean is the best linear unbiased estimator. We demonstrate now its inability to deal with outliers. To show this, it suffices to take any random sample from the normal distribution. We choose one observation from the sample and alter its value. If the value of the observation is far enough from the expected value of the rest of the observations then the mean of the random sample is also shifted. It can even happen that the mean is outside the cloud of not perturbed observations. The most common way to deal with this problem is to replace the mean by median (there are also other possible estimators).

Our sample contains n observations. The computational complexity of its mean is then n and of the median is $n \log(n)$.

The situation for regression does not differ too much. We have a random sample of n observations which follow the model (1.1). As before we choose one observation i. Let the observation follow a different model $y_i = x_i^\top \gamma + \varepsilon_i$, where x_i is a vector of regressors and $\gamma \neq \beta$. Let further x_i lie far outside the rest of regressors i.e., $\max_{k,l \in 1,\dots,i-1,i+1,\dots,n} \|x_k - x_l\| \ll \min_{l \in 1,\dots,i-1,i+1,\dots,n} \|x_i - x_l\|$. If we employ the least squares then a final estimate $\hat{\beta}$ of β can be shifted significantly from β. In this more general case, we mention three robust methods dealing with regression but there are plenty of others.

- The least weighted squares: in this method the weighted sum of squared residuals is minimized with respect to β. We order the statistics of square residuals and put the higher weight to the observations with the lower value of the square residuals. The idea of implicit weighting residuals was proposed by (Víšek, 2000). This method leads to a very robust estimator and has satisfactory properties, see (Mašíček, 2004). Nevertheless, its computation is intensive and an approximative algorithm must be used already for moderate sample sizes, see (Kalina, 2009).

- Regression quantiles: the method generalizes the procedure of minimization of residuals in the absolute norm. We mention it also in Chapter 2. It was introduced in (Koenker and Bassett, 1978). The regression quantiles can be computed with the help of the simplex algorithm (Jurečková and Picek, 2005).
- Theil-Sen algorithm: The idea for two dimensional case is to choose the median of slopes of the lines which connect two points from the sample. It was introduced in (Theil, 1950) and extended in (Sen, 1968). Its properties are still studied (Hanxiang et al., 2008). Several different methods are known for computing the Theil–Sen estimator exactly in $n \log n$ time see (Brönnimann and Chazelle, 1998). We utilize especially the variation proposed by (Siegel, 1982), which is also known as repeated median.

The question arises: How to compare estimators with respect to their robustness? There is no generally accepted definition of robustness. There are plenty of different measures trying to deal with that. We mention two of them.

- Influence function measures the dependence of the estimator on the value of one of the points from the sample (see (v. Mises, 1947)). I.e., we replace the i-th observation of the sample by an arbitrary value and look how the output of the estimator changes. There are some desirable properties of the estimator with respect to its influence function.

  1. Finite rejection point - this means that there is a threshold. If the value of the threshold is exceeded by the replaced observation then the value of the influence function is equal to zero.
  2. Gross error sensitivity should be small. The gross error sensitivity is the supreme absolute value of influence function over all possible values of the replaced observation.
  3. Local shift sensitivity should be small. The local shift sensitivity represents the effect of a small shift, of the replaced observation, from its initial position.

  We present two examples: mean has its gross error sensitivity equal to infinity, local shift sensitivity equal to 1 and infinite rejection point. We compare it with the median, which has a gross error sensitivity equal to $\sqrt{\frac{\pi}{2}}$, local shift sensitivity equal to infinity and finite rejection point (see (Jurečková and Picek, 2005)).

- One of the most important concepts in robust statistics is a breakdown point which was introduced for the first time in (Hampel, 1971) and then was studied in (Donoho and Huber, 1983). Loosely speaking, it expresses a fraction of data which can be "arbitrarily increased" without affecting the finiteness of an estimator. Thus, the high breakdown point is the valuable property of the estimator. However, computational complexity is usually higher for estimators with a high value of the breakdown point and further increases with a dimension of the observations. For the mean the break down point equals to zero and for the median one half constituting the best value for "reasonable" estimators. As an example of the improper estimator let us take one which is always equal to zero. The estimator cannot be violated by any perturbed observation thus its break down point is equal to one.

In our work we focus our attention on this measure of robustness.

Throughout the work we utilize a median as a basic estimator in our methods. We have mentioned that the median can serve as a more robust estimator than the mean and therefore it is natural to replace the mean by the median whenever we wish to gain a more robust method. As it has also been mentioned, the median is employed in repeated median algorithm for linear regression.

The idea of the median is quite old. The concept of it appeared in the Talmud, the book from the 13^{th} century. Pierre Simon Laplace in 1774 suggested the minimization in the absolute norm leading to the median instead of a quadratic norm. The quadratic norm leads to the mean and was suggested by Carl Friedrich Gauss and Adrien Maria Legendre (see (Stigler, 1973)). Francis Galton used the term median in the year 1881 in his work Report of the Anthropometric Committee.

So far we have discussed especially the median appropriate for a one dimensional case, but we want to deal also with general normed vector space and regression. For any normed vector space we utilize, as a natural generalization of median, the concept of geometric median. In the case of the regression we employ the repeated median estimator.

As it has been already mentioned, on the one hand the robust methods handle better violations from the assumptions, on the other hand their computational complexity is in comparison to the classical methods rather poor. In our work we focus not only on robustness but our aim is also to find as simple statistical methods as possible. The methods should also be easy to implement and have low computational complexity. For this purpose the simple idea hidden behind the median serves as a really good starting point.

In our work we present different simulation studies. They are all based on the principle that the normal distribution with predefined parameter of location is perturbed by some other heavy tailed distribution in a selected proportion. We believe that this technique approximates well the real data. After that we compare different estimators with respect to the average absolute deviation from the parameter of location of the not perturbed normal distribution.

In Chapter 2 we focus on recursive adaptive methods which deal with smoothing and forecasting. At first we describe the exponential smoothing. The literature dealing with its robustification (proposing similar method which is not so prone to outliers) remains rather lacking. The first attempt to fulfill this gap was made in (Cipra, 1992) (this method is described later in our work). Other authors have tried to apply robust versions of the Kalman filter to the state-space model associated with exponential smoothing. They generally employ M-estimation. This method is described in (Cipra and Romera, 1992), (Romera and Cipra, 1995), (Gelper et al., 2010) and (Hanzák and Cipra, 2011).

We generalize previous results from (Cipra, 1992) for exponential smoothing in the absolute norm by using the regression quantiles (Koenker and Bassett, 1978). We examine the breakdown point of the method. We also mention possible extensions for more parameters. A method based on a classical sign test algorithm is introduced. It deals not only with outliers but also with level shifts, including a detection of change points. This method is also extended to the more dimensional case. The break down point, complexity, convergence and equivariance is evaluated by the studied algorithms. The ability of estimating is investigated for various approaches by means of a simulation study. We consider not only the ability of the investigated methods to smooth the series but also to forecast. We illustrate the methods on real data example.

We propose new robust estimators for parameter of location in Chapter 3 . A location family of a distribution is a class of probability distributions where the distribution depends on a location parameter a which determines the shift of the distribution. Having a random variable with density $f(x)$, the density of all other distribution in the location family can be expressed as $f_a(x) = f(x - a)$. The classical example of a distribution which depends on the location parameter is the normal distribution.

We follow the same goal of finding an easily computable robust estimator with the high breakdown point in the chapter. For this purpose we employ not only the idea of the median but also the trimmed mean.

Let us describe now the procedure from Chapter 3. At first we try to find the unperturbed observations in our sample. These observations are utilized in the construction of robust estimators. To gain them, we need initially some computationally simple estimator with the high breakdown point. Since the geometric median satisfies these requirements, we employ it as the initial estimator. The geometric median is a direct generalization of the median. It was firstly mentioned in (Weber, 1909) and rediscovered by (Haldane, 1948) and its properties were studied in details by (Kemperman, 1987). It is well defined even if the random variable does not have a finite first moment and it is considered robust because its breakdown point is equal to $\frac{1}{2}$. Even though it is not affine invariant, it is still translation equivariant and scale invariant. In a functional context, consistent estimators of the geometric median were proposed by (Kemperman, 1987), (Cadre, 2001) and (Gervini, 2008). Another advantage of the geometric median is its low computational complexity, which can be easily seen when the absolute norm is employed, as the computation of the geometric median reduces to computation of componentwise median. A vast number of algorithms were proposed for its computation, see (Cardot et al., 2013).

Methods generalizing median were also proposed for time series analysis but only in special case of AR(1) (Zielinski, 1999) and (Luger, 2006). However, the further research in this area would be possible.

Further, we find some set $\mathcal{L}$ from a sample of observations which are robust in the sense of breakdown point around the geometric median. Then we utilize these observations to construct further estimators, for example by enlarging $\mathcal{L}$ and computing the (iterative) weighted mean of observations from the enlarged set. Since the geometric median has the breakdown point of $\frac{1}{2}$, our estimators will be able to keep this property as well.

We try to keep the chapter as general as possible. For this reason, we deal with normed spaces meaning that our estimators are not restricted only on $\mathbb{R}^d$ (where d is a dimension of the space), but we can for instance engage with spaces of integrable functions such as L^q. This opens a natural way to handle time series by our approach. It is also possible to employ our approach in regression, where we could once more utilize the Siegel algorithm instead of the geometric median. Our estimators enjoy the following desirable properties:

- Instead of considering the $\mathbb{R}^d$ space, we work with a general normed vector space X. This opens a natural way to tackle time series by our approach.
- Our estimators have the high breakdown point and are simple to compute.
- We partially consider the covariance structure.
- We are able to work with set-valued estimators instead of single-valued estimators.
- Their computational complexity is low.

Note that estimators usually only satisfy several of the above properties, for example either they have the high breakdown point or they do not take into account the covariance structure at all.

We deal with the consistency of our estimators in a special case of $\mathbb{R}^d$.

We also study how our method can be visualized by implementing it into boxplot and bagplot in the chapter .

1.1 Basic definitions

We introduce some of the basic definitions and notation utilized throughout our work in this part. Since the first part of it deals with time series and the second with parameters of location, our notation slightly differs between these parts. We always highlight the differences.

In this work we employ a general probability space $(\Omega, \mathcal{A}, \mathrm{P})$.

Our notation is basically standard: by $(X, \|\cdot\|)$ we understand a normed vector space. For a set $A \subset X$, we define

$$\|A\| := \sup_{x\in A} \|x\|.$$

For $A, B \subset X$ and $c \in \mathbb{R}$ we define the Minkowski sum, Minkowski difference and multiplication by a scalar as

$$\begin{aligned} A + B &:= \{a + b |\ a \in A,\ b \in B\}, \\ A - B &:= \{a - b |\ a \in A,\ b \in B\}, \\ cA &:= \{ca |\ a \in A\}, \end{aligned}$$

respectively.

We often use the bold notation for $\boldsymbol{x} = (x_1, \ldots, x_n) \in X^n$ and by lower index we understand a component of a vector. In Chapter 2 we employ the bold notation whenever $\boldsymbol{x} \in \mathbb{R}^d$ for $d \in \mathbb{N}$ and $d > 1$.

A multifunction $R : X \rightrightarrows Y$ is a generalization of a function, where the image does not have to be one point but may be a subset of Y.

We also need some basic functions. Let $x \in \mathbb{R}$ and $x \geq 0$ then $\lceil x \rceil$ is the ceiling function. I.e., $\lceil x \rceil = \min\{n \in \mathbb{N} | n \geq x\}$.

We define the floor function as $\lfloor x \rfloor = \max\{n \in \mathbb{N} | n \leq x\}$.

We define the signum function for $x \in \mathbb{R}$ as

$$\operatorname{sgn}(x) = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ -1 & \text{if } x < 0. \end{cases}$$

We put $[x]^+ = \max(0, x)$ and $[x]^- = -\min(0, x)$ for $x \in \mathbb{R}$. This denotes the positive (respectively the negative) part of any real number.

Let f denote one of the previously mentioned functions (ceiling, floor, signum function and positive resp. negative parts) and $\boldsymbol{x} \in \mathbb{R}^d$, $\boldsymbol{x} = (x_1, \dots, x_d)^\top$. We put $f(\boldsymbol{x}) = (f(x_1), \dots, f(x_d))^\top$.

Binomial coefficients $\binom{n}{k}$ for $n, k = 0, 1, 2, \dots$ are defined in a usual way. If $n = 0$ then $\binom{n}{k} = 0$.

Definition 1.1.1. *Consider an estimator $T_n : X^n \rightrightarrows X$. We say that T_n is*

- *shift equivariant if for all $x_1, \dots, x_n \in X$ and $y \in X$ we have*

$$T_n(x_1 + y, \dots, x_n + y) = T_n(x_1, \dots, x_n) + y.$$

- *shift invariant if for all $x_1, \dots, x_n \in X$ and $y \in X$ we have*

$$T_n(x_1 + y, \dots, x_n + y) = T_n(x_1, \dots, x_n).$$

- *scale equivariant if for all $x_1, \dots, x_n \in X$ and $c \in \mathbb{R}$*

$$T_n(cx_1, \dots, cx_n) = cT_n(x_1, \dots, x_n).$$

Definition 1.1.2. *We say that random variable Y has the distribution F which is contaminated by distribution G with probability p if $\mathrm{P}(Y \sim F) = 1 - p$ and $\mathrm{P}(Y \sim G) = p$.*

Similarly, we consider the observations with distribution G as outliers.

We deal with the consistency of estimators and convergence in our work, therefore we have a reminder of some basic definitions.

Definition 1.1.3. *Let $X_1, \dots, X_n$ be a sequence of i.i.d. random variables drawn from a distribution with parameter θ and $T_n(X_1, \dots, X_n)$ an estimator of θ. We say that $T_n(X_1, \dots, X_n)$ is a consistent estimator of θ if for any $\varepsilon > 0$*

$$\lim_{n \to \infty} \mathrm{P}\left(|T_n(X_1, \dots, X_n) - \theta| > \varepsilon\right) = 0.$$

We are sometimes able to show an even stronger kind of convergence.

Definition 1.1.4. *Let $X_1, \dots, X_n$ be a sequence of i.i.d. random variables drawn from a distribution with parameter θ and $T_n(X_1, \dots, X_n)$ an estimator of θ. We say that $T_n(X_1, \dots, X_n)$ is a strongly consistent estimator of θ or that $T_n(X_1, \dots, X_n)$ converges almost surely to θ if for any $\varepsilon > 0$*

$$\mathrm{P}\left(\liminf_{n \to \infty} \left(\omega \in \Omega : |T_n(X_1(\omega), \dots, X_n(\omega)) - \theta| < \varepsilon\right)\right) = 1.$$

From the almost sure convergence follows the convergence in probability.

The Fisher consistence is also mentioned in our work.

Definition 1.1.5. *Let $X, X_1, \ldots, X_n$ be a sequence of i.i.d. random variables drawn from a distribution with parameter θ, $T_n(X_1, \ldots, X_n)$ an estimator of θ, F cumulative distribution function of X and F_n empirical distribution function from the sample $X_1, \ldots, X_n$. If an estimator of θ based on the sample can be represented as a functional of the empirical distribution function, i.e., $T_n(X_1, \ldots, X_n) = T(F_n)$ then we say that estimator is Fisher consistent if $T(F) = \theta$.*

This definition does not have to describe asymptotic properties of the estimator. E.g., we want to estimate expected value μ of some distribution. We utilize X_1 as an estimate of μ regardless of n. This estimator is Fisher consistent; nevertheless, it is not consistent.

A breakdown point is nowadays one of the standard measures of robustness and expresses the minimal proportion of the data which can be corrupted (made arbitrarily distant) before the estimator becomes unbounded. Our definition differs from the one in (Donoho and Huber, 1983), but maintains original properties.

Definition 1.1.6. *Consider a normed vector space X and an estimator $T_n : X^n \rightrightarrows X$ of some functional T. For $\boldsymbol{x} = (x_1, \ldots, x_n) \in X^n$ and $m = 1, \ldots, n$ we define*

$$A_{m,n}(\boldsymbol{x}) := \{\tilde{\boldsymbol{x}} \in X^n \mid \tilde{\boldsymbol{x}} \textit{ and } \boldsymbol{x} \textit{ have at most } m \textit{ different coordinates}\},$$

$$m_n^*(T_n, \boldsymbol{x}) := \max_{m \in \{1,\ldots,n\}} \left\{ m \;\middle|\; \sup_{\tilde{\boldsymbol{x}} \in A_{m,n}(\boldsymbol{x}),\ \tilde{z} \in T_n(\tilde{\boldsymbol{x}}),\ z \in T_n(\boldsymbol{x})} \|\tilde{z} - z\| < \infty \right\}.$$

Then we say that T_n has the breakdown point

$$\varepsilon_n^*(T_n, \boldsymbol{x}) := \frac{1}{n} m_n^*(T_n, \boldsymbol{x}).$$

Finally, for family of estimators $\{T_n : X^n \rightrightarrows X\}$ we define the asymptotic breakdown point as

$$\varepsilon^* := \liminf_{n \to \infty, \boldsymbol{x} \in X^n} \varepsilon_n^*(T_n, \boldsymbol{x}).$$

We denote the observations from $\tilde{\boldsymbol{x}}$ as perturbed if they differ from observations from $\boldsymbol{x}$.

The different approach to the break down point, which also takes into account time series, was studied in (Genton and Lucas, 2003). We deal with time series in Chapter 2, but we consider there recursive adaptive algorithms, for which the method from (Genton and Lucas, 2003) is not suitable.

When working with a time series, it is often crucial to smooth it. To do so, one often considers only neighboring measurements and smooth on their basis. However, if we apply Definition 1.1.6 to the smoothing operator, the resulting breakdown point would be extremely small and would depend on the length of the series. For this reason we propose a modification.

Let $T_n : \mathbb{R}^n \to \mathbb{R}^n$ be an estimator of observations in a time series and $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ is the time series. For simplicity we denote for $t = 1, \ldots, n$ by $T_{n,t} : \mathbb{R}^{m(t,\boldsymbol{y})} \to \mathbb{R}$ the smoothed values (estimates) of y_t. Here, $m(t, \boldsymbol{y})$ denotes the size of the neighborhood $\boldsymbol{z}_t \subset \boldsymbol{y}$ on the base of which the estimate is computed.

We neglect the fact that T_n is usually a map between $\mathbb{R}^n$ and $\mathbb{R}^{n-W}$ for some $W \in \mathbb{N}, W < n$ and not $\mathbb{R}^n \to \mathbb{R}^n$.

Definition 1.1.7. *Let $T_n : \mathbb{R}^n \to \mathbb{R}^n$ be an estimator of time series, $\boldsymbol{y} = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$ and $\tau \leq n$. Then we denote breakdown point of T_n by*

$$\varepsilon_{\tau,n}(T_n, \boldsymbol{y}) := \min_{t=\tau,\dots,n} \varepsilon^*_{m(t,\boldsymbol{y})}(T_{n,t}, \boldsymbol{z}_t),$$

where we utilize Definition 1.1.6 on the right side of the equation.

Consider that we have infinite number of observations $\boldsymbol{y} = (y_1, y_2, \dots)^\top$ and that $T_n \to \bar{T}$ for $n \to \infty$ then the breakdown point for infinite series is defined as

$$\varepsilon(\bar{T}, \boldsymbol{y}) := \liminf_{t \to \infty, \boldsymbol{z}_t \in \mathbb{R}^{m(t,\boldsymbol{y})}} \varepsilon^*_{m(t,\boldsymbol{y})}(\bar{T}_t, \boldsymbol{z}_t),$$

where $\bar{T}_t : \mathbb{R}^{m(t,\boldsymbol{y})} \to \mathbb{R}$ is an estimator of the t^{th} observation.

In some situations is $m(t, \boldsymbol{y})$ a random variable. In this case we consider the worst possible alternative for computing breakdown point.

Definition 1.1.8. *Let $T_n : \mathbb{R}^n \to \mathbb{R}^n$ be an estimator of time series, $\boldsymbol{y} = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$, $\tau \leq n$, $S(t, \boldsymbol{y})$ denotes a set containing all possible $\boldsymbol{z} \subseteq \boldsymbol{y}$ from which can be final estimate $T_{n,t}$ computed, $m(t, \boldsymbol{z})$ denotes the number of observations in $\boldsymbol{z}$. Then we denote breakdown point of T_n by*

$$\varepsilon_{\tau,n}(T_n, \boldsymbol{y}) := \min_{t=\tau,\dots,n; \boldsymbol{z} \in S(t,\boldsymbol{y})} \varepsilon^*_{m(t,\boldsymbol{z})}(T_{n,t}, \boldsymbol{z}),$$

where we utilize Definition 1.1.6 on the right side of the equation.

Consider that we have an infinite number of observations $\boldsymbol{y} = (y_1, y_2, \dots)^\top$ and that $T_n \to \bar{T}$ for $n \to \infty$ then the breakdown point for infinite series is defined as

$$\varepsilon(\bar{T}, \boldsymbol{y}) := \liminf_{t \to \infty, \boldsymbol{z} \in S(t,\boldsymbol{y})} \varepsilon^*_{m(t,\boldsymbol{z})}(\bar{T}_t, \boldsymbol{z}),$$

where we utilize Definition 1.1.6 on the right side of the equation.

We employ Definition 1.1.8 in Part 2.2.1.

Example 1.1.9. *Let us have time series $\boldsymbol{y} = (y_1, \dots, y_n)^\top$ which follows the model*

$$y_t = a_t + \varepsilon_t \quad \text{for } t = 1, \dots, n,$$

where ε_t is a white noise process. As a norm we employ the Euclidian norm $\|\cdot\|_2$.

Let $W \in \mathbb{N}$ and $W \geq 2$. We denote the weighted median from observations $y_t, \dots, y_{t-W+1}$ by $M(y_t, \dots, y_{t-W+1})$ where the observation y_t has a weight $\frac{2}{W+1}$ and y_i has a weight $\frac{1}{W+1}$ for $i = t-1, \dots, t-W+1$. Our algorithm has the form.

Algorithm 1.1.1 Smoothing with a weighted median

```
Input: observations $(y_1, \dots, y_n)$, $W \geq 2$ and $W \leq n$
compute $M(y_W, \dots, y_1)$
put $\hat{y}_W \leftarrow M(y_W, \dots, y_1)$
put $t \leftarrow W + 1$
while $t \leq n$ do
    $\hat{y}_t \leftarrow M(y_t, \dots, y_{t-W+1})$
    $t \leftarrow t + 1$
end while
```

We employ Algorithm 1.1.1 as a functional T_n from Definition 1.1.7 . We put $\tau = W$. In our case $m(t, \boldsymbol{y}) = W$ and $\boldsymbol{z}_t = (y_{t-W+1}, \dots, y_t)^\top$. Due to the properties of the weighted median we get $\varepsilon^*_W \left(T_{n,t}, (y_{t-W+1}, \cdots, y_t)^\top\right) = \frac{1}{W}\left\lfloor \frac{W-2}{2} \right\rfloor$. From that we get $\varepsilon_{\tau,n}(T_n, \boldsymbol{y}) = \frac{1}{W}\left\lfloor \frac{W-2}{2} \right\rfloor$ and therefore also $\varepsilon(\bar{T}, \boldsymbol{y}) = \frac{1}{W}\left\lfloor \frac{W-2}{2} \right\rfloor$ for $n \to \infty$. If we utilize a simple median, we will get $\varepsilon_{\tau,n}(T_n, \boldsymbol{y}) = \frac{1}{W}\left\lfloor \frac{W-1}{2} \right\rfloor$ instead. △

We should note that Definition 1.1.6 of the asymptotic breakdown point differs from the breakdown point for the infinite series from 1.1.7, because $m(t, \boldsymbol{y})$ does not generally converge to infinity in the case of 1.1.7 when $n \to \infty$.

We can consider a different definition of the breakdown point for time series as a stricter alternative . Namely, we put $\boldsymbol{y}$ instead of $\boldsymbol{z}_t$ and n instead of $m(t, \boldsymbol{y})$ in Definition 1.1.7. Then, if we perturb observations $\boldsymbol{z}_t$ and not the rest of the observations, we could get a really low break down point. This complies with the direct application of Definition 1.1.6.

In the case of time series we normally do not employ all observations to smooth the time series in the time t. The position of contaminated observations is very important in this case. E.g., if the perturbed observations lie in the beginning of the sample then the smoothed values of the series in the time t which is also in the beginning can also be wrong.

Let us suppose that we know the probability p that the observation y_t is perturbed. We are interested in the probability that the smoothed value of the time series in the time t is not perturbed.

Definition 1.1.10. *Let us have $\boldsymbol{y} = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$ and $T_n : \mathbb{R}^n \to \mathbb{R}^n$ an estimator of time series. Further, $p = \mathrm{P}(\bar{y}_t = y^*_t)$ and $1 - p = \mathrm{P}(\bar{y}_t = y_t)$ for $t = 1, \dots, n$, where y^*_t is any number from $\mathbb{R}$, $p \in [0, 1]$ and $\bar{\boldsymbol{y}} = (\bar{y}_1, \dots, \bar{y}_n)^\top$. Let further the indicators $I(\bar{y}_1 = y^*_1), \dots, I(\bar{y}_n = y^*_n)$ be independent. We define the probability $\pi_t(T_n)$ of unviolated solution as*

$$\pi_t(T_n) = \mathrm{P}\left(\sup_{(y^*_1,\dots,y^*_n)\in\mathbb{R}^n} \|T_{n,t}(\boldsymbol{y}) - T_{n,t}(\bar{\boldsymbol{y}})\| < \infty \right).$$

Proposition 1.1.11. *Let us have $\boldsymbol{y} = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$ and $T_n : \mathbb{R}^n \to \mathbb{R}^n$ an estimator of time series. The observation y_t for $t = 1, \dots, n$ can be perturbed with probability $p \in [0, 1]$ in the sense of Definition 1.1.10. Further, we employ $\tau \in \mathbb{N}$ such that $\tau \le n$. Then for $t \ge \tau$*

$$\pi_t(T_n(\bar{\boldsymbol{y}})) \ge \sum_{i=0}^{\lfloor \varepsilon_{\tau,n}(T_n,\bar{\boldsymbol{y}}) m(t,\bar{\boldsymbol{y}}) \rfloor} p^i (1-p)^{m(t,\bar{\boldsymbol{y}})-i} \binom{m(t,\bar{\boldsymbol{y}})}{i}. \tag{1.2}$$

Proof. If the number of perturbed observations exceed $\lfloor \varepsilon_{\tau,n}(T_n, \bar{\boldsymbol{y}}) m(t, \bar{\boldsymbol{y}}) \rfloor$ then we get from Definition 1.1.7 that $\hat{y}_t(\boldsymbol{z}_t)$ can be influenced by the observations from $(y^*_1, \dots, y^*_n)^\top$. For the indicators $I(\bar{y}_1 = y^*_1), \dots, I(\bar{y}_n = y^*_n)$ are independent we utilize binomial distribution with parameters $m(t, \bar{\boldsymbol{y}})$ and p. □

The inequality in (1.2) stands for the fact that $\varepsilon_{\tau,n}(T_n, \bar{\boldsymbol{y}}) \le \varepsilon^*_{m(t,\boldsymbol{y})}(T_{n,t}, \boldsymbol{z}_t)$ for any $t = \tau, \cdots, n$ and that the observations from $\boldsymbol{z}_t$ can influence $\hat{\boldsymbol{y}}_t(\boldsymbol{z}_t)$ differently.

Example 1.1.12. *The assumptions in this example are the same as in Example 1.1.9. We utilize the simple exponential smoothing in the following form to smooth the series $\boldsymbol{y}$.*

Algorithm 1.1.2 Simple exponential smoothing

```
Input: observations $(y_1, \dots, y_n)$, $W \in \mathbb{N}$, $W < n$ and $0 < \lambda < 1$
1: compute mean $\mu_W$ from $(y_1, \dots, y_W)^\top$
2: put $\hat{y}_W \leftarrow \mu_W$
3: put $t \leftarrow W + 1$
4: while $t \leq n$ do
5:     $\hat{y}_t \leftarrow \lambda y_t + (1 - \lambda)\hat{y}_{t-1}$
6:     $t \leftarrow t + 1$
7: end while
```

We employ Algorithm 1.1.2 as a functional T_n from Definition 1.1.7. We put $\tau = W$. In this case $m(t, \boldsymbol{y}) = t$. For some $i \in \mathbb{N}$ satisfying $i \leq n$ and $t \in \mathbb{N}$ such that $n \geq t \geq i$ we denote $\bar{\boldsymbol{y}}_{i,t} = (y_1, \dots, y_{i-1}, y_i^*, y_{i+1}, \dots, y_t)^\top$. If $y_i^* \to \infty$ then from Steps 2 and 5 follows $T_{n,t}\left(\bar{\boldsymbol{y}}_{i,t}\right) \to \infty$, which yields $\|T_{n,t}(\bar{\boldsymbol{y}}_{i,n}) - T_{n,t}(\boldsymbol{y})\|_2 \to \infty$. Since $\bar{\boldsymbol{y}}_{i,t} \subset A_{1,t}(y_1, \dots, y_t)$ (see Definition 1.1.6), we get $\varepsilon_t^*\left(T_{n,t}, (y_1, \dots, y_t)^\top\right) = 0$, which yields $\varepsilon_{\tau,n}(T_n, \boldsymbol{y}) = 0$ and finally $\varepsilon(\bar{T}, \boldsymbol{y}) = 0$ for $n \to \infty$.

We fix $t \geq W$ and compute $\pi_t(T_n(\bar{\boldsymbol{y}}))$. From $\varepsilon_t^*\left(T_{n,t}, (y_1, \dots, y_t)^\top\right) = 0$ for $t = 1, \dots, n$ we get $\pi_t(T_n(\bar{\boldsymbol{y}})) = (1-p)^t$. △

Example 1.1.9. *We continue with Example 1.1.9.*

We compute $\pi_t(T_n(\bar{\boldsymbol{y}}))$ for fixed $t \geq W$. We get

$$\pi_t(T_n(\bar{\boldsymbol{y}})) \geq \sum_{i=0}^{\left\lfloor \frac{W-2}{2} \right\rfloor} p^i (1-p)^{W-i} \binom{W}{i}$$

from Proposition 1.1.11.

Nevertheless, $\pi_t(T_n(\bar{\boldsymbol{y}}))$ is higher, for there is a mismatch in weights of $y_t, \dots, y_{t-W+1}$. We have to consider separately the cases when y_t is contaminated and when it is not. We get

$$\pi_t(T_n(\bar{\boldsymbol{y}})) = p \sum_{i=0}^{\left\lfloor \frac{W-2}{2} \right\rfloor - 1} p^i (1-p)^{W-1-i} \binom{W-1}{i} + \\ + (1-p) \sum_{i=0}^{\left\lfloor \frac{W}{2} \right\rfloor} p^i (1-p)^{W-1-i} \binom{W-1}{i}.$$

△

We will also deal with a time complexity of our algorithms. We employ the basic definitions from (Sipser, 2012). We omit a precise definition of Turing machine. It can be found in (Sipser, 2012, p. 167). The following description can serve as a vague idea of it. An infinite tape serves as unlimited memory of the Turing machine model. It has a tape head that can read and write symbols and move around on the tape. Initially the tape contains only an input string and is blank everywhere else. If the machine needs to store information, it may write it on the tape. To read it the machine can move its head back over it. The machine computes until it decides to produce an output.

We adopt the following definitions from (Sipser, 2012, p. 276, 277).

Definition 1.1.13. *Let M be a deterministic Turing machine that halts on all inputs. The running time or time complexity of M is the function $f : \mathbb{N} \to \mathbb{N}$, where $f(n)$ is the maximum number of steps that M uses on any input of length n.*

The exact running time is usually quite a complex expression. Therefore, we consider only the highest order term of the expression for the running time of the algorithm, disregarding both the coefficient of that term and any lower order terms, because the highest order term dominates the other terms on large inputs.

These reasons we lead us to the following notation. We write $f(n) = \mathcal{O}(g(n))$ for functions $f, g : \mathbb{N} \to \mathbb{R}^+$ if for some $c \in \mathbb{R}$, exists n_0 such that for all $n \geq n_0$, $f(n) \leq cg(n)$.

Example 1.1.12. *To compute the running time of Algorithm 1.1.2 we only need to sum two numbers for each t. Therefore, the time complexity is equal to $\mathcal{O}(n)$.*

Example 1.1.9. *We discuss the time complexity of Algorithm 1.1.1. We have to compute the weighted median from W observations during the algorithm . The most demanding part of it is to order the observations. To do so, we utilize a heapsort algorithm. Its time complexity is $\mathcal{O}(W \log W)$ according to (Cormen, 2009)[151]. However, we have to order the observations just once. In other words, we can exploit the ordering for t from $t-1$ and therefore, we have for each t the time complexity only $\mathcal{O}(W)$. It yields the time complexity $\mathcal{O}(Wn)$, but W can be considered as a constant with respect to n and therefore the final complexity is just $\mathcal{O}(n)$.*

△

2. Recursive adaptive methods

We deal especially with methods very similar to exponential smoothing in the beginning of this chapter. I.e., the methods which employ exponential weights. Further, we suppose an alternative approach, which is based on the idea of the sign test.

The simplicity and recursive computing scheme predetermine the exponential smoothing to be widely used for time series. It is employed for smoothing and forecasting. It is an ad hoc procedure, but there are connections to ARIMA models, see (Brown, 1962).

It is shown in (Papageorgiou et al., 2005) that the method is still effective in practical problems.

However, the exponential smoothing, as many other statistical methods, is very sensitive to outliers. We have mentioned in the introduction that some authors attempt to deal with this problem by introducing the robust version of the Kalman filter. The approach based on it supposes that there can be a change in level in each step, even if the change is rather small. On the other hand we suppose, that level shifts appear in our data only rarely, but the corresponding changes can be really significant then.

We attempt to employ L-estimators here, see (Jurečková and Picek, 2005), and we test for level shifts.

The general exponential smoothing supposes the model of the form

$$y_t = \boldsymbol{z}_t^\top \boldsymbol{a}_t + \varepsilon_t, \tag{2.1}$$

where $\{y_t\}_{t=1}^n$ is a given time series, $\boldsymbol{a}_t$ vector of parameters, $\boldsymbol{z}_t$ vector of fitting functions (both of these vectors are of dimension d), and ε_t a white noise. The white noise is usually supposed to be i.i.d. with normal distribution. We loosen these assumptions and suppose ε_t to be i.i.d. with the median equal to zero. We usually employ normally distributed ε_t contaminated by distributions with heavy tails but with a probability density symmetric around the origin. Compare with Definition 1.1.2.

The classical approach of exponential smoothing operates in the l_2 norm (see, e.g., (Hyndman et al., 2008)). One looks for adaptive estimates $\hat{\boldsymbol{a}}_t((y_1, \dots, y_t), \beta)$ by minimizing

$$\sum_{i=1}^{t} \beta^{t-i} (y_i - \boldsymbol{z}_i^\top \boldsymbol{a}_t)^2 \tag{2.2}$$

at time t, where $\beta \in (0, 1)$ is a discount coefficient. The solution is unique except for degenerated cases.

Two robust approaches replacing general exponential smoothing are considered in this chapter:

1. the exponential smoothing using regression quantiles and implemented by means of a special algorithm in l_1 norm,

2. the approach combining some ideas of the exponential smoothing with the classical sign test. This method can also be applied to time series with level shifts.

2.1 Exponential smoothing based on regression quantiles

The objective function (2.2) to be minimized can, using the methodology of the regression quantiles with a robustifying effect (see (Koenker and Bassett, 1978)), be transformed to the form

$$\sum_{i=1}^{t} \beta^{t-i} \varrho_\alpha(y_i - \boldsymbol{z}_i^\top \boldsymbol{a}_t), \tag{2.3}$$

where $\alpha \in (0,1)$ and

$$\varrho_\alpha(x) = |x|\{\alpha I[x \geq 0] + (1-\alpha) I[x < 0]\}, \quad x \in \mathbb{R}.$$

For a given α, we denote one of the solutions of (2.3), by $\boldsymbol{a}_t^\alpha((y_1,\ldots,y_t),\beta)$. Then the corresponding smoothed value of $\alpha-$quantile of y_t is defined as $y_t^\alpha = (\boldsymbol{z}_t)^\top \boldsymbol{a}_t^\alpha((y_1,\ldots,y_t),\beta)$. It generalizes the l_1 approach (Cipra, 1992). Indeed, in the case of the median, i.e., with $\alpha = 0.5$, instead of (2.3) we solve the minimization problem

$$\sum_{i=1}^{t} \beta^{t-i} |y_i - \boldsymbol{z}_i^\top \boldsymbol{a}_t| \tag{2.4}$$

(see, e.g., (Cipra, 1992), (Cipra and Romera, 1992) and (Romera and Cipra, 1995)). The approach based on the regression quantiles can follow certain ideas used in the previous works with $\alpha = 0.5$.

We approximate (2.3) by

$$\sum_{i=t-W+1}^{t} \beta^{t-i} \varrho_\alpha(y_i - \boldsymbol{z}_i^\top \boldsymbol{a}_t) \tag{2.5}$$

to solve it, where W should be large enough that the observations
$y_{t-W}, y_{t-W-1}, \ldots, y_1$ with weights $\beta^W, \beta^{W+1}, \ldots, \beta^{t-1}$ can be neglected.

The minimization of function (2.5) is discussed in the following Section.

2.1.1 Regression quantiles: Algorithm for the generalized l_1 approach

We introduce an algorithm for solving (2.5) generalizing the l_1 approach to the exponential smoothing here. We describe the case $d = 1$ first. We deal with the general multi-dimensional case then.

2.1.2 Case $d = 1$

We employ a simple algorithm for the case of $d = 1$. This algorithm has been introduced by (Cipra, 1992) for $\alpha = 0.5$, and we refer to it as C-algorithm.

We look for

$$\tilde{a}_t^\alpha((y_1,\ldots,y_n),W,\beta) \in \underset{a\in\mathbb{R}}{\operatorname{argmin}} \left\{ \sum_{i=t-W+1}^{t} \beta^{t-i} \varrho_\alpha(y_i - z_i a) \right\}, \tag{2.6}$$

where z_i are known scalars. We omit $((y_1, \ldots, y_n), W, \beta)$, to avoid burdensome notation, when we write $\tilde{a}_t^{\alpha}((y_1, \ldots, y_n), W, \beta)$. An equality is not expressed in Formula (2.6), because its solution is not generally unique. We can assume without loss of generality that $z_1, \ldots, z_t \neq 0$. In particular, it includes the case of the classical constant trend for time series for $z_i \equiv 1$ and $\alpha = 0.5$ and one obtains a robust version of simple exponential smoothing. The algorithm has the following form for general z_i :

Algorithm 2.1.1 Exponential smoothing in l_1 norm case $d = 1$

```
Input: $(y_1, \ldots, y_n)^\top$, $(z_1, \ldots, z_n)^\top$, $W \in \mathbb{N}$, $\beta, \alpha \in (0, 1)$
$t \leftarrow W$
while $t \leq n$ do
    order the ratios $\frac{y_{t-W+1}}{z_{t-W+1}}, \ldots, \frac{y_t}{z_t}$
    denote the ordered values by $v_{(1)} \leq v_{(2)} \leq \cdots \leq v_{(W)}$
    for $j = t - W + 1, \ldots, t$ do
        find $i$ such that $v_{(i)} = \frac{y_j}{z_j}$
        if $z_j > 0$ then
            $c_i^- \leftarrow \alpha\beta^{t-j}|z_j|$ and $c_i^+ \leftarrow (1-\alpha)\beta^{t-j}|z_j|$
        else
            $c_i^- \leftarrow (1-\alpha)\beta^{t-j}|z_j|$ and $c_i^+ \leftarrow \alpha\beta^{t-j}|z_j|$
        end if
    end for
    find the index $r$ $(r = \{1, \ldots, W\})$ which fulfills

        $\sum_{j=1}^{r-1} c_j^+ - \sum_{j=r}^{W} c_j^- < 0$
        $\sum_{j=1}^{r} c_j^+ - \sum_{j=r+1}^{W} c_j^- \geq 0$        (2.7)

    put $\tilde{a}_t^{\alpha} \leftarrow v_{(r)}$
end while
```

We construct the quantile estimate in this way .

The idea behind Algorithm 2.1.1 is as follows: The minimum of the objective function (2.6) can be a point or an interval; nevertheless, there are no other local minima. I.e., the function descends from minus infinity then it attains the minimum, whether it is a point or an interval, and then it ascends into infinity. The algorithm looks for one of the points, where the derivative of the objective function changes its sign. The derivative is not defined in the points $v_{(1)}, \ldots, v_{(W)}$.

The procedure can be performed recursively if one exploits the ordering of $v_{(1)}, \ldots, v_{(W)}$ from the previous step. We have to put current ratio $\frac{y_t}{z_t}$ into the proper place and get rid of $\frac{y_{t-W}}{z_{t-W}}$. We can order $v_{(1)}, \ldots, v_{(W)}$ by the heapsort or another sorting algorithm with a low computational complexity in the beginning.

Lemma 2.1.1. *Let $T_n : \mathbb{R}^n \to \mathbb{R}^n$ be an estimator defined by Algorithm 2.1.1 with inputs $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, $z_t = 1$ for $t = 1, \ldots, n$, then $\tilde{a}_t^{\alpha}$ is for $t = W, \ldots, n$ a shift equivariant estimate. If further $\alpha = 0.5$ then $\tilde{a}_t^{\alpha}$ is for $t = W, \ldots, n$ also scale equivariant.*

Proof. Let all observations be shifted by $s \in \mathbb{R}$. Then for $t = W, \dots, n$

$$\sum_{i=t-W+1}^{t} \beta^{t-i} \varrho_\alpha(y_i + s - a_t) = \sum_{i=t-W+1}^{t} \beta^{t-i} \varrho_\alpha(y_i - (a_t - s)). \tag{2.8}$$

I.e., if the solution of (2.6) is $\tilde{a}_t^\alpha((y_1, \dots, y_n), W, \beta)$ then for the minimum $\tilde{a}_t^\alpha((y_1 + s, \dots, y_n + s), W, \beta)$ of (2.8) holds

$$\tilde{a}_t^\alpha((y_1 + s, \dots, y_n + s), W, \beta) - s = \tilde{a}_t^\alpha((y_1, \dots, y_n), W, \beta).$$

If $\alpha = 0.5$ and $c, x \in \mathbb{R}$ then $\varrho_{0.5}(cx) = c\varrho_{0.5}(x)$. This yields for $t = W, \dots, n$ and $c \neq 0$

$$\sum_{i=t-W+1}^{t} \beta^{t-i} \varrho_{0.5}(cy_i - a_t) = c \sum_{i=t-W+1}^{t} \beta^{t-i} \varrho_{0.5}(y_i - \frac{1}{c} a_t). \tag{2.9}$$

From (2.9) we get

$$\frac{1}{c} \tilde{a}_t^{0.5}((cy_1, \dots, cy_n), W, \beta) = \tilde{a}_t^{0.5}((y_1, \dots, y_n), W, \beta).$$

The case $c = 0$ is simple. □

Example 2.1.2. *We show simple examples, when the Algorithm 2.1.1 is not shift and scale equivariant. Let* $W = 2$, $y_1 = -2$, $y_2 = 1$, $z_1 = 1$, $z_2 = 2$, $\beta = 1$ *and* $\alpha = 0.5$. *The solution of* (2.6) *is* $\frac{1}{2}$. *If we plug into* (2.6) $y_1 + s$ *and* $y_2 + s$, *where* $s = 2$, *we get the solution* $\frac{3}{2}$.

Let now $\alpha = 0.9$, $z_2 = 1$ *and the rest of the values stay the same. Then the solution of* (2.6) *is* 1. *If we plug into* (2.6) cy_1 *and* cy_2, *where* $c = -1$, *we get the solution* 2. △

If we employ the series $-y_1, \dots, -y_n$ instead of $y_1, \dots, y_n$ in Algorithm 2.1.1 with $c < 0$ and $\alpha \neq 0.5$ we should get $\tilde{a}_t^\alpha((-y_1, \dots, -y_n), W, \beta) = -\tilde{a}_t^{1-\alpha}((y_1, \dots, y_n), W, \beta)$. Therefore, the fact that the algorithm is not scale equivariant for $\alpha \neq 0.5$ is reasonable.

Further, we deal with the convergence of C-algorithm. It cannot converge in the case that we employ any window W. We simplify our task for a while such that we are looking for a in a model

$$y_t = a + \varepsilon_t, \tag{2.10}$$

where we suppose ε_t to be i.i.d. non-degenerated random variables with the median equal to 0. The constant a does not change over time. We are looking for a by minimizing

$$\sum_{i=1}^{t} \beta^{t-i} |y_i - a|, \tag{2.11}$$

for each $t = 1, 2, \dots$. If $\beta = 1$ then the almost sure convergence to the median is proved e.g., in (Pollard, 2012). In the case that $0 < \beta < 1$ we show that because of its exponential weights there appears the same problem as for exponential smoothing, namely that the weights of more recent observations are too high.

Proposition 2.1.3. *Let us have a time series following the model* (2.10) *and estimates $\hat{a}_t$ gained as the solution of* (2.11). *Then $\hat{a}_t$ does not converge to a almost surely neither in probability.*

Proof. We fix some $t \in \mathbb{N}$. We want to compute the least number k of the most recent observations $y_t, y_{t-1}, \ldots, y_{t-k+1}$ of which the sum of weights is always greater than one half of the sum of all weights. The sum of all possible weights is

$$\sum_{i=0}^{\infty} \beta^i = \frac{1}{1-\beta}.$$

The sum of weights of the most recent k observations is

$$\sum_{i=0}^{k-1} \beta^i = \frac{1-\beta^k}{1-\beta}.$$

We are looking for the least k for which

$$\frac{1}{2(1-\beta)} \leq \frac{1-\beta^k}{1-\beta}.$$

This yields

$$\frac{1}{2} \leq 1 - \beta^k.$$

There is such a number k because $0 < \beta < 1$ and therefore $\lim_{k\to\infty} 1 - \beta^k = 1$.

Since ε_t is non-degenerated it holds $\mathrm{P}(\varepsilon_t > c > 0) = \delta > 0$ or $\mathrm{P}(\varepsilon_t < c < 0) = \delta > 0$. We suppose without loss of generality that $\mathrm{P}(\varepsilon_t > c > 0) = \delta > 0$. It yields that there is nonzero probability that k consecutive y_t are greater than a. If we split the series $\{y_t\}_{t=1}^{\infty}$ into the distinct parts of the length k such that $t_1 = 1, t_2 = k+1, \ldots$ and denote by E_{t_j} the event that $y_{t_j} > a+c, \ldots, y_{t_j+k} > a+c$ then we get $\sum_{j=1}^{\infty} \mathrm{P}(E_{t_j}) = \infty$. The events E_{t_j} are independent, therefore we can employ Borell-Cantelli lemma. It yields that E_{t_j} occurs for infinitely many j almost surely. We can split the series $\{y_t\}_{t=1}^{\infty}$ in any way i.e., that for any $t \in \mathbb{N}$ the event that $y_t > a, \ldots, y_{t+k} > a$ can occur and these events arise for infinitely many t.

If E_{t+1} occurs then, because of the choice of k, the sum of weights corresponding to observations $y_{t+1}, \ldots, y_{t+k}$ is greater than the sum of the rest of the weights. It yields that $\mathrm{P}(\tilde{a}_{t+k} > a + c \text{ i.o.}) = 1$. It also shows that $\tilde{a}_t$ cannot converge in probability. □

We show a result relating to the robustness of the C-algorithm now. We employ Definition 1.1.7 of the breakdown point in the following lemma.

Lemma 2.1.4. *Let $T_n : \mathbb{R}^n \to \mathbb{R}^n$ be an estimator defined by Algorithm 2.1.1 with inputs $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, $z_t = 1$ for $t = 1, \ldots, n$, $\alpha = 0.5$, $\beta \in (0,1)$ and $W \in \mathbb{N}$, $W \leq n$. Further, define*

$$j = \max_{i=1,\ldots,T} \left\{ i \,\middle|\, \frac{1-\beta^{i-1}}{1-\beta^{T-1}} < \frac{1}{2} \right\}, \tag{2.12}$$

then the breakdown point of T_n for $W, \ldots, n$ is given by the ratio $\frac{j}{W}$.

Proof. We utilize Definition 1.1.7 in this proof. $m(t, \boldsymbol{y}) = W$ for any $t \in \{W, \dots, n\}$. We are therefore investigating the breakdown point of $T_{n,t,W}(y_{t-W+1}, \dots, y_t)$ according to Definition 1.1.6. We want to find an estimate $T_{n,t,W}(y_{t-W+1}, \dots, y_t)$ for which the minimum of the sum

$$\sum_{i=t-W+1}^{t} \beta^{t-i} |y_i - a_t| \tag{2.13}$$

is attained. It gives us a weighted median, but the weights β^{t-i} are not normalized. To normalize them we have to sum all weights from (2.13), which gives $\frac{1-\beta^W}{1-\beta}$. The weights of the nearest observations to the t are the highest. Therefore, they influence $T_{n,t,W}(y_{t-W+1}, \dots, y_t)$ the most. Let us find j such that the sum of the normalized weights of observations $y_{t-j+1}, \dots, y_t$ is greater than $\frac{1}{2}$. It follows from the fact, that we deal with a weighted median, that by moving the $y_{t-j+1}, \dots, y_t$ observations far enough we can also move arbitrarily with $T_{n,t,W}(y_{t-W+1}, \dots, y_t)$. We can see On the other side that if the sum of normalized weights of $y_{t-j+1}, \dots, y_t$ is less than $\frac{1}{2}$ then the movement of these observations does not influence the estimate.

I.e., to compute the breakdown point we have to find maximal j for which the sum of normalized weights of the observations $y_{t-j+1}, \dots, y_t$ is less than $\frac{1}{2}$. But this is given by (2.12). □

To evaluate the robustness in the sense of Definition 1.1.10 we can simply employ Proposition 1.1.11. The equality does not hold, for the observations have different weights in (1.2).

The weights of observations $y_{t-W+1}, \dots, y_t$ are $\frac{\beta^{W-1}(1-\beta)}{1-\beta^W}, \dots, \frac{1-\beta}{1-\beta^W}$ for the step t of Algorithm 2.1.1. We denote the weight of observation y_i in the step t as $w(t, y_i)$. It holds $w(t, y_i) = 0$ for $i \leq t - W$. By $\Pi(t, W, \beta)$ we denote the set of combinations from the indices $t, \dots, t - W + 1$ such that

$$\Pi(t, W, \beta) = \left\{ \kappa \,\middle|\, \sum_{i \in \kappa} w(t, y_i) < \frac{1}{2} \right\}.$$

Let $\kappa \in \Pi(t, W, \beta)$, $T_n : \mathbb{R}^n \to \mathbb{R}^n$ be an estimator defined by Algorithm 2.1.1 with proper inputs and we denote the number of indices in κ by $\#\kappa$. We get

$$\pi_t(T_n(\bar{\boldsymbol{y}})) = \sum_{\kappa \in \Pi(t,W,\beta)} p^{\#\kappa}(1-p)^{W-\#\kappa}.$$

We want to discuss a time complexity (see Definition 1.1.13) of Algorithm 2.1.1 now. It resembles Example 1.1.9. We have to recompute for each t median from W observations, but we can exploit the ordering from the previous step. Therefore, we have complexity $\mathcal{O}(Wn)$ or more roughly $\mathcal{O}(n)$.

2.1.3 General case

The idea behind the algorithm in the general case i.e., with $d > 1$ is the same as before. I.e., to find a point around which the signs of all directional derivatives of the function (2.5) change (in the case that the minimum of the function (2.5) is uniquely defined).

The algorithm described in this section is not implemented and we also do not deal with its theoretical properties for its computational complexity and difficult implementation, which is given by a more complex principle of ordering in higher dimensions. We can also find in literature some suitable alternatives. Nevertheless, to preserve the generality of our work we will also discuss this topic. We only mention some heuristics connecting to this topic and the ideas behind them from these reasons.

We put

$$g_t(\boldsymbol{x}) = \sum_{i=t-W+1}^{t} \beta^{t-i} \varrho_\alpha(y_i - \boldsymbol{z}_i^\top \boldsymbol{x}),$$

where we for simplicity neglect variables $W, \alpha, \beta, (y_1, \ldots, y_n)^\top, (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^\top$ on which $g_t(\boldsymbol{x})$ also depends to emphasize our interest in t and $\boldsymbol{x}$.

We want to minimize the objective function $g_t(\boldsymbol{x})$, which is convex, because its components are convex. Therefore, we are seeking the point $\hat{\boldsymbol{a}}_{\boldsymbol{t}}$ satisfying for any $\boldsymbol{v} \in \mathbb{R}^d$ and $h > 0$ one of the following

$$\begin{aligned} &\operatorname{sgn}(g_t(\hat{\boldsymbol{a}}_{\boldsymbol{t}} + h\boldsymbol{v}) - g_t(\hat{\boldsymbol{a}}_{\boldsymbol{t}})) = -\operatorname{sgn}(g_t(\hat{\boldsymbol{a}}_{\boldsymbol{t}} - h\boldsymbol{v}) - g_t(\hat{\boldsymbol{a}}_{\boldsymbol{t}})) \text{ or} \\ &\quad g_t(\hat{\boldsymbol{a}}_{\boldsymbol{t}} + h\boldsymbol{v}) - g_t(\hat{\boldsymbol{a}}_{\boldsymbol{t}}) = 0 \text{ or} \\ &\quad g_t(\hat{\boldsymbol{a}}_{\boldsymbol{t}} - h\boldsymbol{v}) - g_t(\hat{\boldsymbol{a}}_{\boldsymbol{t}}) = 0. \end{aligned} \tag{2.14}$$

We are not employing directional derivatives now, because they are not defined for all $\boldsymbol{x} \in \mathbb{R}^d$.

We suppose in the following text.

- We employ only W of the most recent observations as in the case $d = 1$.
- We suppose that there is always at least one point lying on the intersection of d different hyperplanes.
- We suppose that $W \geq d$.

Let $\boldsymbol{z}_i \in \mathbb{R}^d$ be a known vector for some $i = t - W + 1, \ldots, t$. The least value of $|y_i - \boldsymbol{z}_i^\top \boldsymbol{x}|$ is attained for $\boldsymbol{x}$ for which $y_i - \boldsymbol{z}_i^\top \boldsymbol{x} = 0$. We know that the gradient of $g_t(\boldsymbol{x})$ changes around points $\boldsymbol{x}$ for which $y_i - \boldsymbol{z}_i^\top \boldsymbol{x} = 0$ from the properties of absolute value. Therefore, we restrict to $\boldsymbol{x}$ satisfying $y_i - \boldsymbol{z}_i^\top \boldsymbol{x} = 0$ when seeking the minimum of $g_t(\boldsymbol{x})$. The equation $y_i - \boldsymbol{z}_i^\top \boldsymbol{x} = 0$ represents a hyperplane in $\mathbb{R}^d$ with respect to $\boldsymbol{x}$. We can conjecture that the minimum is attained in the intersection of these hyperplanes, because in these intersections more members of $g_t(\boldsymbol{x})$ are minimized. Hence, we have to look for the minimum of $g_t(\boldsymbol{x})$ in the intersection of the hyperplanes, i.e., usually at a point which is given by d different nonparallel hyperplanes. To denote hyperplanes we employ the same index as the index of relevant member of $g_t(\boldsymbol{x})$. I.e., the corresponding hyperplane has an index equal to i and we denote it as H_i for the summand $\varrho_\alpha(y_i - \boldsymbol{z}_i^\top \boldsymbol{x})$ of $g_t(\boldsymbol{x})$.

We summarize these findings into two steps.

1. Find all points given by an intersection of d hyperplanes from the set $H_{t-W+1}, \ldots, H_t$. Denote the set Π_t.

2. A point from Π_t fulfilling (2.14) minimizes $g_t(\boldsymbol{x})$.

This minimization problem can be solved with the help of a special table, where any row represents a line given by the intersection of appropriate $d-1$ hyperplanes. Let us denote the table B_t. There are points given by an intersection of the line and another hyperplane in each column. This table is updated for each t. We can employ a similar algorithm as in the one-dimensional case after the construction of the table to find the minimum.

We deal with a construction of the table and evaluation of conditions (2.14) in the following text.

Structure of the table B_t

We describe how the previously mentioned table B_t should be constructed in this part.

Each row expresses an intersection of $d-1$ different hyperplanes. Because we always take a different combination of the hyperplanes, the row is uniquely defined by the hyperlanes. Therefore, the combination identifies the row. The intersection of the hyperplanes represents a line in $\mathbb{R}^d$. Such a table has $\binom{W}{d-1}$ rows.

The labels change for each time and the table has to be recomputed.

An example of the identifiers of the rows for time t is shown in Table 2.1. Each row is identified by specific indices of the hyperplanes.

Table 2.1. There is shown the ordering of rows in the tableB_t.

$t-W+d-1, t-W+d-2,$ $t-W+d-3, \dots, t-W+1$	$\cdot$	$\cdot$	$\cdots$	$\cdot$
$t-W+d, t-W+d-2,$ $t-W+d-3, \dots, t-W+1$	$\cdot$	$\cdot$	$\cdots$	$\cdot$
$t-W+d, t-W+d-1,$ $t-W+d-3, \dots, t-W+1$	$\cdot$	$\cdot$	$\cdots$	$\cdot$
		$\vdots$		
$t-W+d, t-W+d-1,$ $t-W+d-2, \dots, t-W+2$	$\cdot$	$\cdot$	$\cdots$	$\cdot$
		$\vdots$		
$t, t-W+d-2,$ $t-W+d-3, \dots, t-W+1$	$\cdot$	$\cdot$	$\cdots$	$\cdot$
		$\vdots$		
$t, t-1, t-2, \dots, t-W+1$	$\cdot$	$\cdot$	$\cdots$	$\cdot$
		$\vdots$		
$t, t-1, t-2, \dots, t-d+2$	$\cdot$	$\cdot$	$\cdots$	$\cdot$

Consider I_t the set of indices of all hyperplanes $\{t-W+1, \dots, t\}$. We denote the set of indices given in the r^{th} row of the table B_t by $I_{r,t}$, where $r = 1, \dots, \binom{W}{d-1}$ e.g., $I_{1,t} = \{t-W+d-1, t-W+d-2, t-W+d-3, \dots, t-W+1\}$.

Information in fields of the table B_t

We have dealt only with the first column of B_t in the previous part. We want to show now what we store in its fields.

Each row of the table B_t expresses a line in R^d. Consider $r = 1, \dots, \binom{W}{d-1}$ then a field of the row r contain information about a point given by intersection of hyperplanes with indices $I_{r,t}$ and a hyperplane with index from $I_t \setminus I_{r,t}$. There are $W - d + 1$ such fields in each row.

We store information about the coordinates of the corresponding point in each field.

Remark 2.1.5. *Since each point is given by d equations we can for the calculation of coordinates employ Gaussian elimination. However, for higher dimensions than two we can simplify the whole procedure. E.g., we adjust the equations given by hyperplanes with indices $I_{r,t}$ by Gaussian elimination for the r^{th} row of B_t. Then we add an equation given by the hyperplane with index $j \in I_t \setminus I_{r,t}$, which relates to the considered field.*

Since each row expresses a line in $\mathbb{R}^d$ and its field the point on that line, we order the fields of the row according to its coordinates. It is possible to order the points in each line according to only one coordinate. Therefore, it is usually possible to order the points according to the first coordinate. If this is not the case, then we employ the second, the third, or the $d-^{\text{th}}$ coordinate.

Let us consider now any field of the table B_t in the r^{th} row containing a point $\boldsymbol{x} \in \Pi_t$. We know that the point is given by the intersection of hyperplanes with indices $I_{r,t}$ and one more hyperplane with index $j_1 \in I_t \setminus I_{r,t}$. We want to gather information about the indices of hyperplanes on which $\boldsymbol{x}$ is lying. We denote the set of these indices $P_t(\boldsymbol{x})$. We get $I_{r,t} \cup j_1 \subseteq P_t(\boldsymbol{x})$. However, we can find other hyperplanes with indeces $j_2, \dots, j_l$ during the ordering such that they also lie on the intersection of hyperplanes with indices $I_{r,t} \cup j_k$, where $k = 2, \dots, l$. It yields $P_t(\boldsymbol{x}) = I_{r,t} \cup j_1 \cup j_2 \cup \dots \cup j_l$.

We want to store information helping us to evaluate the conditions (2.14). Since the gradient of $g_t(\boldsymbol{x})$ is not defined at $\boldsymbol{x}$ and we need information about change of slope in $\boldsymbol{x}$, we apply the following procedure. We define

$$f_t(\boldsymbol{x}) = \sum_{i=t-W+1}^{t} \beta^{t-i} \varrho_\alpha(y_i - \boldsymbol{z}_i^\top \boldsymbol{x}) - \sum_{i \in P_t(\boldsymbol{x})} \beta^{t-i} \varrho_\alpha(y_i - \boldsymbol{z}_i^\top \boldsymbol{x}).$$

We store the vector $\boldsymbol{b}_t(\boldsymbol{x}) = \left[\frac{\partial f_t(\boldsymbol{a})}{\partial a_1}, \dots, \frac{\partial f_t(\boldsymbol{a})}{\partial a_d}\right]^{\boldsymbol{a}=\boldsymbol{x}}$ in each field of B_t, where $\boldsymbol{a} \in \mathbb{R}^d$.

We introduce vectors

$$\boldsymbol{c}_{i+,t} = \beta^{t-i} \alpha \boldsymbol{z}_i, \quad \boldsymbol{c}_{i-,t} = \beta^{t-i}(1-\alpha)\boldsymbol{z}_i. \tag{2.15}$$

Let $P_{+,t}(\boldsymbol{x}) \subset I_t$ be a set of indices such that for all $i \in P_{+,t}(\boldsymbol{x})$ holds $y_i - \boldsymbol{z}_i^\top \boldsymbol{x} > 0$. Similarly, we define $P_{-,t}(\boldsymbol{x}) \subset I_t$ such that for all $i \in P_{-,t}(\boldsymbol{x})$ holds $y_i - \boldsymbol{z}_i^\top \boldsymbol{x} < 0$. We can write, employing the notation, $\boldsymbol{b}_t(\boldsymbol{x}) = \sum_{i \in P_{+,t}(\boldsymbol{x})} \boldsymbol{c}_{i+,t} - \sum_{i \in P_{-,t}(\boldsymbol{x})} \boldsymbol{c}_{i-,t}$.

We have to deal also with the hyperplanes of which indices belong to $P_t(\boldsymbol{x})$ to observe the change of gradient of $g_t(\boldsymbol{x})$ around $\boldsymbol{x}$. Therefore, we store the following vectors

$$\begin{aligned} \boldsymbol{b}_{+,t}(\boldsymbol{x}) &= \sum_{i \in P_t(\boldsymbol{x})} [\boldsymbol{c}_{i+,t}]^+ + \sum_{i \in P_t(\boldsymbol{x})} [\boldsymbol{c}_{i-,t}]^+, \\ \boldsymbol{b}_{-,t}(\boldsymbol{x}) &= \sum_{i \in P_t(\boldsymbol{x})} [\boldsymbol{c}_{i+,t}]^- + \sum_{i \in P_t(\boldsymbol{x})} [\boldsymbol{c}_{i-,t}]^-. \end{aligned} \tag{2.16}$$

Conditions for global minimum

We show the conditions of $\boldsymbol{x}$ being the global minimum of $g(\boldsymbol{x})$ in this part. They replace conditions (2.14).

Let $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{x}$ is one of the points from the table B_t. If it satisfies

$$\operatorname{sgn}(\boldsymbol{b}_t(\boldsymbol{x}) + \boldsymbol{b}_{+,t}(\boldsymbol{x})) = -\operatorname{sgn}(\boldsymbol{b}_t(\boldsymbol{x}) - \boldsymbol{b}_{-,t}(\boldsymbol{x})), \tag{2.17}$$

and every component of vectors $\boldsymbol{b}_t(\boldsymbol{x}) + \boldsymbol{b}_{+,t}(\boldsymbol{x})$ and $\boldsymbol{b}_t(\boldsymbol{x}) - \boldsymbol{b}_{-,t}(\boldsymbol{x})$ is different from zero then the point $\boldsymbol{x}$ is a sharp global minimum of the function $g(\boldsymbol{x})$. In the case that one component of the vector $\boldsymbol{b}_t(\boldsymbol{x}) + \boldsymbol{b}_{+,t}(\boldsymbol{x})$ or $\boldsymbol{b}_t(\boldsymbol{x}) - \boldsymbol{b}_{-,t}(\boldsymbol{x})$ is equal to zero, then there is a global minimum but not necessarily sharp. We denote such a vector $\boldsymbol{x}$ as $\tilde{\boldsymbol{a}}_t^\alpha$ i.e., the solution of our problem.

The intuition behind (2.17) comes from the fact that absolute value function changes its sign in the minimum.

We split the description of the algorithm into two parts. We deal with the initiation of the table B_W in the first part and with a general step t in the second part.

Construction of B_t

Consider some row r of B_t and some index $j \in I_t \setminus I_{r,t}$. If the hyperplanes with indices $I_{r,t} \cup j$ do not generate $\mathbb{R}^d$ then according to the Frobenius theorem there is no point on the intersection of these hyperplanes.

Therefore, we cannot place any information to the appropriate field of B_t.

It may also happen that the hyperplanes given by indices $I_{r,t}$ do not generate $\mathbb{R}^{d-1}$. In this case it does not make sense to compute any point in such a row. However, we should place information about it.

The worst case takes place if none from the d-tuples of hyperplanes gives a point. We can consider this case for W high enough really rare. The problem can be fixed by adding some new hyperplanes. Such hyperplane with index i does not have to influence the computation of gradient; therefore, we put $\boldsymbol{c}_{i+,t} = \boldsymbol{c}_{i-,t} = 0$.

We neglect these degenerated cases in a description of the rest of the algorithm.

We stress that if we talk about lines, we mean only lines which are given by the intersection of hyperplanes with indices $I_{r,t}$, where r is an index of the row of the table B_t. Similarly, by talking about a point we mean the point $\boldsymbol{x} \in \Pi_t$.

We mean by preparing the table B_t construction of the first column depicted in Table 2.1.

- We mean the whole construction of the table for $t = d$.
- We mean the addition of appropriate rows containing index t in the first column for $d < t \leq W$. The indices started always from 1 and B_t contains $\binom{t}{d-1}$ rows.

- We mean the elimination of rows containing index $t - W$ and addition of rows with index t in the first column for $t > W$.

Point $\boldsymbol{x}(P)$ denotes the point given by the intersection of hyperplanes with indices from appropriate index set P.

We define $m_{r,t}$ as a number of fields in the row r at time t.

Let $\boldsymbol{x}$ be in B_t in r^{th} line on the o^{th} place after ordering. We define $O(o, r, t) = P_t(\boldsymbol{x})$.

Algorithm 2.1.2 Exponential smoothing in l_1 norm case $d > 1$

```
Input: $(y_1, \dots, y_n)^\top$, $(\boldsymbol{z}_1, \dots, \boldsymbol{z}_n)^\top$, $W \in \mathbb{N}$, $\beta, \alpha \in (0, 1)$
t ← d
prepare $B_t$
compute $\boldsymbol{x}(\{I_{1,t} \cup t\})$                    ▷ there is only one point for all rows
$P_t(\boldsymbol{x}(\{I_{1,t} \cup t\})) \leftarrow \{I_{1,t} t\}$
$\boldsymbol{b}(\boldsymbol{x}(\{I_{1,t} \cup t\})) \leftarrow 0$
for $r = 1, \dots, d$ do
    compute $\boldsymbol{c}_{r+,t}$ and $\boldsymbol{c}_{r-,t}$ according to (2.15)
end for
compute $\boldsymbol{b}_{+,t}(\boldsymbol{x}(\{I_{1,t} \cup t\}))$ and $\boldsymbol{b}_{-,t}(\boldsymbol{x}(\{I_{1,t} \cup t\}))$ according to (2.16)
put the information about $\boldsymbol{x}(\{I_{1,t} \cup t\})$ into rows according to 2.1.6
for $t = d + 1, \dots, W$ do
    add to $B_t$ the rows containing index $t$ according to 2.1
    for $r = 1, \dots, \binom{t-1}{d-1}$ do
        compute $\boldsymbol{x}(\{I_{r,t} \cup t\})$
        put $\boldsymbol{x}(\{I_{r,t} \cup t\})$ on the proper place in the row $r$
        find $P_t(\boldsymbol{x}(\{I_{r,t} \cup t\}))$
        compute $\boldsymbol{c}_{t+,t}$ and $\boldsymbol{c}_{t-,t}$ according to (2.15)      ▷ comment see below
        compute $\boldsymbol{b}_{+,t}(\boldsymbol{x}(\{I_{r,t} \cup t\}))$ and $\boldsymbol{b}_{-,t}(\boldsymbol{x}(\{I_{r,t} \cup t\}))$ according to (2.16)
        for $o = 1, \dots, m_{r,t}$ do
            if $y_t - \boldsymbol{z}_t^\top \boldsymbol{x}(O(o, r, t)) > 0$ then
                $\boldsymbol{b}_{+,t}(\boldsymbol{x}(O(o, r, t))) \leftarrow \beta \boldsymbol{b}_{+,t-1}(\boldsymbol{x}(O(o, r, t)))$
                $\boldsymbol{b}_{-,t}(\boldsymbol{x}(O(o, r, t))) \leftarrow \beta \boldsymbol{b}_{-,t-1}(\boldsymbol{x}(O(o, r, t)))$
                $\boldsymbol{b}(\boldsymbol{x}(O(o, r, t))) \leftarrow \beta \boldsymbol{b}(\boldsymbol{x}(O(o, r, t))) + \boldsymbol{c}_{t+,t}$
            end if
```

```
25:             if $y_t - \boldsymbol{z}_t^\top \boldsymbol{x}(O(o,r,t)) < 0$ then
26:                 $\boldsymbol{b}_{+,t}(\boldsymbol{x}(O(o,r,t))) \leftarrow \beta \boldsymbol{b}_{+,t-1}(\boldsymbol{x}(O(o,r,t)))$
27:                 $\boldsymbol{b}_{-,t}(\boldsymbol{x}(O(o,r,t))) \leftarrow \beta \boldsymbol{b}_{-,t-1}(\boldsymbol{x}(O(o,r,t)))$
28:                 $\boldsymbol{b}(\boldsymbol{x}(O(o,r,t))) \leftarrow \beta \boldsymbol{b}(\boldsymbol{x}(O(o,r,t))) - \boldsymbol{c}_{t-,t}$
29:             end if
30:             if $y_t - \boldsymbol{z}_t^\top \boldsymbol{x}(O(o,r,t)) = 0$ then
31:                 continue with 2.1.3
32:                 put the information $\boldsymbol{x}(\{I_{r,t} \cup t\})$ into rows according to 2.1.6
33:             end if
34:         end for
35:     end for
36: end for
37: return estimate $\hat{x} \leftarrow z^k$
```

We introduce auxiliary index sets I^+ and I^- in the following procedure. We neglect all the variables and parameters on which they depend to simplify the notation.

Algorithm 2.1.3 Procedure for the case $y_t - \boldsymbol{z}_t^\top \boldsymbol{x}(O(o,r,t)) = 0$

```
if $P_t(\boldsymbol{x}(O(o,r,t))) \setminus \{I_{r,t} \cup t\} = \varnothing$ then
    if $o < m_{r,t}$ then
        $I^+ = \{i | i \in P_t(\boldsymbol{x}(O(o+1,r,t))) \setminus \{I_{r,t}, y_i - \boldsymbol{z}_i^\top \boldsymbol{x}(O(o,r,t)) > 0\}$
        $I^- = \{i | i \in P_t(\boldsymbol{x}(O(o+1,r,t))) \setminus \{I_{r,t}, y_i - \boldsymbol{z}_i^\top \boldsymbol{x}(O(o,r,t)) < 0\}$
        $\boldsymbol{b}(\boldsymbol{x}(O(o,r,t))) \leftarrow \beta \boldsymbol{b}(\boldsymbol{x}(O(o+1,r,t))) + \sum_{i \in I^+} \boldsymbol{c}_{i+,t} - \sum_{i \in I^-} \boldsymbol{c}_{i-,t}$
    else
        $I^+ = \{i | i \in P_t(\boldsymbol{x}(O(o-1,r,t))) \setminus \{I_{r,t}, y_i - \boldsymbol{z}_i^\top \boldsymbol{x}(O(o,r,t)) > 0\}$
        $I^- = \{i | i \in P_t(\boldsymbol{x}(O(o-1,r,t))) \setminus \{I_{r,t}, y_i - \boldsymbol{z}_i^\top \boldsymbol{x}(O(o,r,t)) < 0\}$
        if $y_t - \boldsymbol{z}_t^\top \boldsymbol{x}(O(o,r,t)) > 0$ then
            $\boldsymbol{b}(\boldsymbol{x}(O(o,r,t))) \leftarrow \boldsymbol{b}(\boldsymbol{x}(O(o-1,r,t))) + \sum_{i \in I^+} \boldsymbol{c}_{i+,t} - \sum_{i \in I^-} \boldsymbol{c}_{i-,t} - \boldsymbol{c}_{t+,t}$
        else
            $\boldsymbol{b}(\boldsymbol{x}(O(o,r,t))) \leftarrow \boldsymbol{b}(\boldsymbol{x}(O(o-1,r,t))) + \sum_{i \in I^+} \boldsymbol{c}_{i+,t} - \sum_{i \in I^-} \boldsymbol{c}_{i-,t} + \boldsymbol{c}_{t-,t}$
        end if
    end if
else
    $\boldsymbol{b}(\boldsymbol{x}(O(o,r,t))) \leftarrow \beta \boldsymbol{b}(\boldsymbol{x}(O(o,r,t)))$
end if
```

We want to find the index set $P_t(\boldsymbol{x}(\{I_{r,t} \cup t\}))$ in Step 16 of Algorithm 2.1.2. Let us suppose that there was a point $\boldsymbol{x} = \boldsymbol{x}(\{I_{r,t} \cup t\})$ in the r^{th} row and step $t-1$, then $P_t(\boldsymbol{x}(\{I_{r,t} \cup t\})) = P_{t-1}(\boldsymbol{x}) \cup t$.

We recompute $\boldsymbol{c}_{i+,t}$ and $\boldsymbol{c}_{i-,t}$ for all $1 \leq i < t$ by multiplying $\boldsymbol{c}_{i+,t-1}$ and $\boldsymbol{c}_{i+,t-1}$ by β in Step 17 of Algorithm 2.1.2.

We recommend for ordering new hyperplanes to employ heapsort see e.g., (Cormen, 2009)[151].

If we add a new index t to Table 2.1 then we first consider the old lines (without index t). Compare to Algorithm 2.1.2 Step 13. However, all points lying on H_t

are already given by the intersection of these old lines and hyperplane H_t. I.e., information about the points contained in the new lines (with index t) was already derived for the old lines.

The following lemma describes how to find a number of row r according to its indices $I_{r,t}$.

Lemma 2.1.6. *Let $t \leq W$, $P_t(\boldsymbol{x}) = \{i_1, \dots, i_l\}$ and $i_1 < \dots < i_l$, where $\boldsymbol{x}$ is a point given by the intersection of $H_{i_1}, \dots, H_{i_l}$ and i_j are the indices of hyperplanes for $j = 1, \dots, l$. Then we find the point $\boldsymbol{x}$ in table B_t on row r with indices $I_{r,t} = \{i_1, \dots, i_{d-1}\}$ for the first time.*

There are $\binom{l}{d}$ rows in which $\boldsymbol{x}$ lies.

The row with indices $\{i_1, i_2, \dots i_{d-1}\}$ is situated in our table on a position $r = \binom{i_{d-1}-1}{d-1} + \binom{i_{d-2}-1}{d-2} + \binom{i_{d-2}-1}{d-3} + \dots + i_1$. We can specify any number of row r according to its indices $I_{r,t}$ in the same way.

Proof. An index i_{d-1} appears in Table 2.1 after all lower indices for the first time. I.e., the first row with index i_{d-1} appears after all rows given by combinations of lower indices than i_{d-1}. There are $\binom{i_{d-1}-1}{d-1}$ such rows. Due to $i_{d-1} > \dots > i_1$ the row with indices $\{i_1, i_2, \dots i_{d-1}\}$ lies among rows with the highest index i_{d-1}. However, these rows are not split in Table 2.1. If we leave the index i_{d-1} from the description of these rows, it constitutes a subtable fulfilling the same assumptions as the original table only containing $d-2$ indices in the first column. Therefore, we can proceed in a similar way as before and prove that there are $\binom{i_{d-2}-1}{d-2}$ rows in the subtable before the first row containing i_{d-2}.

Mathematical induction gives us that there are $r = \binom{i_{d-1}-1}{d-1} + \binom{i_{d-2}-1}{d-2} + \binom{i_{d-2}-1}{d-3} + \dots + i_1 - 1$ rows before the first row containing all indices $\{i_{d-1}, \dots, i_1\}$. □

As it is visible so far Algorithm 2.1.2 is already quite complicated, which is, as was mentioned in the beginning, in contradiction to the principles of our work. It does not also bring, except for generality, big advantages. Therefore, we skip the description for $t > W$, which is similar to the 2.1.2. However, we have to also delete rows with index $t - W$ and also adjust appropriately the vector $\boldsymbol{b}(\boldsymbol{x})$, where $\boldsymbol{x} \in \Pi_t$.

2.1.4 $\alpha-$Winsorized estimator

It might be suitable not to smooth the entire series but only observations lying far from the rest. In this case we utilize an algorithm similar to $\alpha-$Winsorized estimator. It is the procedure when we replace given parts of a sample at the high and low end with the most extreme remaining values. The result of our algorithm is smoothed series $\tilde{y}_t$. We employ similar notation to Algorithm 2.1.1 with the exception of considering the multidimensional case.

Algorithm 2.1.4 $\alpha-$Winsorized algorithm

```
Input: $(y_1, \dots, y_n)^\top$, $(z_1, \dots, z_n)^\top$, $W \in \mathbb{N}$, $\beta, \alpha \in (0,1)$
for $t = 1, \dots, W$ do
    $\tilde{y}_t \leftarrow y_t$
end for
for $t = W, \dots, n$ do
    compute $\tilde{\boldsymbol{a}}_t^{\alpha}$ and $\tilde{\boldsymbol{a}}_t^{1-\alpha}$
    if $\boldsymbol{z}_t^\top \tilde{\boldsymbol{a}}_t^{\alpha} > y_t$ then
        $\tilde{y}_t \leftarrow \boldsymbol{z}_t^\top \tilde{\boldsymbol{a}}_t^{\alpha}$
    else if $\boldsymbol{z}_t^\top \tilde{\boldsymbol{a}}_t^{1-\alpha} < y_t$ then
        $\tilde{y}_t \leftarrow \boldsymbol{z}_t^\top \tilde{\boldsymbol{a}}_t^{1-\alpha}$
    else
        $\tilde{y}_t \leftarrow y_t$
    end if
end for
```

Remark 2.1.7. *We indicate at the end of this part how the problem of minimization in* (2.5) *is very often dealt with. We want to solve the minimization problem*

$$\begin{array}{rrcl} \min\limits_{u_{t-W+1},\dots,u_t,\boldsymbol{a}_t} & \sum_{i=t-W+1}^{t} \beta^{t-i} u_{i,t} & & \\ s.t. & u_{i,t} & \geq & \varrho_\alpha(y_i - \boldsymbol{z}_i^\top \boldsymbol{a}_t) \textit{ for } i = t-W+1, \dots, t, \\ & u_{i,t} & \leq & \varrho_\alpha(y_i - \boldsymbol{z}_i^\top \boldsymbol{a}_t) \textit{ for } i = t-W+1, \dots, t. \end{array} \tag{2.18}$$

Problem (2.18) *can be solved by any linear programming technique. The computational speed and comparison with the least squared error minimization can be found in (Portnoy and Koenker, 1997).*

2.2 Sign test

An alternative approach to the robustification of the exponential smoothing described in this Section can provide even better results. It combines some ideas also employed for exponential smoothing with the classical sign test, see, e.g., (Moore and Wallis, 1943) and (Wolfowitz, 1944). It seems to be applicable also to data with level shifts, including detection of change points. Since the observations are not exponentially weighted, we cannot say that the method from this Section belongs among exponential smoothing methods; in fact, it is a recursive adaptive method.

The sign test was firstly used in 1710 by John Arbuthnot according to (Sprent and Smeeton, 2001, p. 14). He observed that in each year from 1629 to 1710 the number of males christened in London exceeded the number of females. He considered the result as a strong evidence that the probability of male birth is higher than one half.

A rough idea for a recursive estimate of the parameters $\boldsymbol{a}_t$ from (2.1) can be described as follows:

1. Find a robust estimate of $\boldsymbol{a}_t$ based on the median from a segment of observations from the beginning. In the case that $\boldsymbol{a}_t$ is a vector we have

to construct an auxiliary series for each component of $\boldsymbol{a}_t$ and utilize the median of the auxiliary series as the robust estimate of the component of $\boldsymbol{a}_t$.

2. If too many consequent observations lie under or above the estimate then there could be a level shift. The detected change point for the level shift occurs at the point where this pattern begins.

3. Estimate $\boldsymbol{a}_t$ up to this point and start this procedure again from the identified change point.

The algorithms described in this Section are referred to as *sign test algorithms*. In particular, we distinguish the *constant sign test algorithm* 2.2.1 and the *linear sign test algorithm* 2.2.3.

2.2.1 Sign test: constant trend

Let us assume the simplest version of the model (2.1)

$$y_t = a_t + \varepsilon_t, \tag{2.19}$$

where level shifts can occur. We remind our assumption that ε_t are i.i.d. with its theoretical median equal to 0.

We employ the following notation. The length of the time series $\{y_t\}$ is n. We denote the time in which the j^{th} level shift was found as t_j. Consider the time $t \geq t_j$ and that there is no time t_i such that $t_j < t_i \leq t$ and t_i is the time of another level shift. Moreover, consider a median $M_t(t_j, (y_1, \ldots, y_t))$ form observations $y_{t_j}, \ldots, y_t$. We simplify the notation and write only $M_t(t_j)$ instead of $M_t(t_j, (y_1, \ldots, y_t))$ in this section. We always consider the median from the last found change point t_j in the following text. The symbol $\hat{a}_t$ denotes the estimate of a_t at time t and simultaneously the smoothed value of y_t due to the model (2.19). The smoothed values $\hat{a}_t$ between two neighboring change points are chosen as equal; they are given by the last estimated value $M_{t_{j+1}-1}(t_j)$ before the next change point.

The number of observations for the initial computation of the median in the beginning and after each change point is a fixed number $W \in \mathbb{N}$. The choice of W is discussed in 2.3.1.

The idea of the sign test can be exploited in our recursive algorithm to deal with level shifts.

For $r \in \mathbb{R}$ we define

$$I_t(r) = \begin{cases} 1 & \text{if } y_t > r, \\ \frac{1}{2} & \text{if } y_t = r, \\ 0 & \text{if } y_t < r. \end{cases}$$

Moreover, we consider

$$S_t^{t+k}(r) = \sum_{i=t}^{t+k} I_i(r) \text{ for } k \in \mathbb{N} \text{ and } t > t_j. \tag{2.20}$$

In the following considerations about determination of level shift we neglect the possibility that the observation is exactly identical to the median. It approximately holds

$$\mathrm{P}(y_t > M_t(t_j)) = \frac{1}{2}.$$

If we assume that there is no level shift between observations y_t and y_{t+k}, and $\mathrm{P}(y_i = M_{t_j+i}(t_j)) = 0$ for $i = 0, \ldots, k$ then $S_t^{t+k}(M_{t+k}(t_j))$ has approximately binomial distribution with parameters $k+1$ and $\frac{1}{2}$.

In the case that $S_t^{t+k}(M_{t+k}(t_j))$ is too large or too low, one can claim a conjecture that there is a change point because too many observations lie above or below the median.

Similar to a sign test algorithm, we define for $r \in \mathbb{R}$ the statistics

$$A_t^{t+k}(r) = \frac{2S_t^{t+k}(r) - k - 1}{\sqrt{k+1}}. \tag{2.21}$$

We want to identify the potential level shifts in the series $\{y_t\}_{t=1}^n$ with the help of statistics $A_t^{t+k}(M_{t+k}(t_j))$. Therefore, we employ a symmetrical interval $(-b, b)$, where b is determined empirically (see 2.3.1). If there is any $t > t_j$ such that $A_t^{t+k}(M_{t+k}(t_j))$ lies outside the interval $(-b, b)$ then we indicate the first such t as a level shift, put $t_{j+1} = t$ and employ t_{j+1} as the new starting point. We also put $\hat{a}_t = M_{t_{j+1}-1}(t_j)$ for each $t = t_j, \ldots, t_{j+1} - 1$.

Remark 2.2.1. *The computation of $M_{t+1}(t_j)$ from $M_t(t_j)$ is quite straightforward. We hold a vector of ordered observations $y_{t_j}, \ldots, y_t$ from the step t. We plug the observation y_{t+1} into this vector and find $M_{t+1}(t_j)$. We are sparing the computational time with the help of this recursive technique.*

We suppose to have $M_t(t_j)$ for computation of $A_{t-i}^t(M_t(t_j))$ for all $i = 0, \ldots, t - t_j - 1$. We put $S_t^t(M_t(t_j)) = I_t(M_t(t_j))$. We get $S_{t-i}^t(M_t(t_j)) = S_{t-i+1}^t(M_t(t_j)) + I_{t-i}(M_t(t_j))$ according to (2.20). We gain all $S_{t-i}^t(M_t(t_j))$ for $i = 0, \ldots, t - t_j - 1$ in this manner. We derive $A_{t-i}^t(M_t(t_j))$ for all $i = 0, \ldots, t - t_j - 1$ with (2.21).

Lemma 2.2.2. *Consider the case with a real change point at time $\tau \in \mathbb{N}$, $\tau > t_j$ and the constant a_t from (2.19) changing from the value a^o to a^n. In the case that $y_{\tau-1} > M_t(t_j)$ and $A_\tau^t(M_t(t_j)) > b$ for some $t \in \mathbb{N}$ and $t > \tau$ then also $A_{\tau-1}^t(M_t(t_j)) > b$. This arises especially in the case $a^n > a^o$.*

In the opposite case when $y_{\tau-1} < M_t(t_j)$ and $A_\tau^t(M_t(t_j)) < -b$ then $A_{\tau-1}^t(M_t(t_j)) < -b$. This arises especially in the case $a^n < a^o$.

Proof. We show, without loss of generality, only the first case.

It holds for $y_{\tau-1} > M_t(t_j)$ $S_{\tau-1}^{\tau-1+k}(M_t(t_j)) = S_\tau^{\tau-1+k}(M_t(t_j)) + 1$ and therefore also $A_{\tau-1}^{t-\tau+1}(M_t(t_j)) > A_\tau^{t-\tau+1}(M_t(t_j)) > b$. □

The previous lemma shows that we can misspecify the time of the level shift. We suppose to deal with this problem, we have found level shift at time $\tau - 1$ and for the absolute errors from (2.19) holds $\mathrm{E}\,|\varepsilon_{\tau-1}| \ll |a^o - a^n|$. If $a^n > a^o$, $y_{\tau-1} > a^o$ and $|a^o - y_{\tau-1}| < |a^n - y_{\tau-1}|$, we conjecture that there is no change point at time $\tau - 1$. We employ the same procedure to examine whether it is at time τ. We utilize $M_{\tau-1}(t_j)$ and $M_{\tau-1+W}(\tau - 1)$ as estimates of a^o and a^n.

This modification is useful if the errors of the time series do not exceed the difference between a^o and a^n. It is implemented in our algorithm 2.2.1 in Steps 11 - 15.

We summarize our algorithm as follows:

Algorithm 2.2.1 Algorithm for a constant trend based on the sign test

```
Input: observations $(y_1, \dots, y_n)^\top$, $W \in \mathbb{N}$, $b > 0$
 1: put $t \leftarrow 1$, $j \leftarrow 1$, $t_1 \leftarrow 1$
 2: while $t \leq n$ do            ▷ this has to be checked throughout the While loop
 3:     $t \leftarrow t_j + W - 1$
 4:     repeat
 5:         $t \leftarrow t + 1$
 6:         compute $M_t(t_j)$
 7:         compute $A^t_{t_j+k}(M_t(t_j))$ for $k = t - t_j, \dots, 1$
 8:     until $A^t_{t_j+k}(M_t(t_j)) \in (-b, b)$ for $k = t - t_j, \dots, 1$
 9:     $j \leftarrow j + 1$
10:     $t_j \leftarrow t_{j-1} + \min \left\{ k \middle| \left| A^t_{t_{j-1}+k}(M_t(t_{j-1})) \right| > b; k = t - t_{j-1}, \dots, 1 \right\}$
11:     compute $M_{t_j+W}(t_j)$
12:     while $|M_{t_j+W}(t_j) - y_t| > |M_{t_j-1}(t_{j-1}) - y_t|$ and $t_j < t$ do
13:         $t_j \leftarrow t_j + 1$
14:         compute $M_{t_j+W}(t_j)$
15:     end while
16:     $\hat{a}_{t_{j-1}}, \dots, \hat{a}_{t_j-1} \leftarrow M_{t_j-1}(t_{j-1})$
17: end while
```

We can examine the distribution of $A^{t+k}_t(M_{t+k}(t_j))$. We can approximate it by normal distribution for large values of k, but it is generally not appropriate, for k does not have to be sufficiently large. We can exploit the binomial nature of $S^{t+k}_t(M_{t+k}(t_j))$. Nevertheless, this solution would be more computationally demanding, therefore we are satisfied with the empirical choice of b, which is but quite crude. Another disadvantage of employing normal quantiles for b arises from the dependence among the statistics $A^{t+k}_t(M_{t+k}(t_j))$ for different k.

A different approach was studied e.g. in (Koubková, 2004) where the statistics based on residuals defined as a difference between the observation and median were suggested.

We can employ CUSUM tests (Page, 1954) instead of statistics $A^{t+k}_t(r)$. It is a sequential analysis technique which serves in our case for detecting a change in the data. The CUSUM test for median, therefore suitable in our situation, was described in (Yang et al., 2010). The problem of this approach is higher computational complexity and more problematic determination of the observation where the trend has changed.

We can also employ some other nonparametric tests instead of the sign test, e.g., the Wilcoxon signed-rank test (see (Wilcoxon, 1945)). Then we proceed as follows. We compute $M_t(t_j)$. The test investigates whether the location parameter of observations $y_{t-k}, \dots, y_t$ is equal to $M_t(t_j)$ for some $k \in \mathbb{N}$ and $k \leq t - t_j$. Put $Y_i = y_i - M_t(t_j)$ for $i = t - k, \dots, t$. Order $|Y_{t_j}|, \dots, |Y_t|$ and let R_i be the order of $|Y_i|$. Put $S^+ = \sum_{Y_i \geq 0, i=t-k,\dots,t} R_i$ and $S^- = \sum_{Y_i < 0, i=t-k,\dots,t} R_i$. If

$\min(S^+, S^-)$ is too small within the framework of the Wilcoxon test, we have found a change point. This test has to be performed for all $k = 1, \ldots, t - t_j$. Obviously, the observations have to be ordered in each test of this type, which is more time-consuming than our procedure.

Generally, the literature dealing with change point or change detection is quite vast see e.g., (Polunchenko and Tartakovsky, 2012). In this sense, there is a potential to replace sign test algorithm with another method and consider whether the replacement leads to an improvement. We need the tests concerning sequential change point detection. We can employ procedures based on M-estimation e.g., (Hušková, 2014) or (Koubková, 2006). We can also employ rank statistics see (Hušková and Sen, 1989). We choose the sign test algorithm because of its simplicity, robust properties and sequential setup and it is the first option concerning the topic of sequential change point detection in our case. However, there is a space for utilizing other methods.

Lemma 2.2.3. *Let us have observations $(y_1, \ldots, y_n)^\top$, $W \in \mathbb{N}$ and $b > 0$. We define $\hat{a}_t$ according to Algorithm 2.2.1, then $\hat{a}_t$ for $t = 1, \ldots, n$ is shift and scale equivariant.*

Proof. Median is shift and scale equivariant. Let $t, t_j \in \{1, \ldots, n\}$ and $t_j < t$. If $y_t < M_t(t_j, (y_1, \ldots, y_t))$ then also for $c > 0$ $cy_t < M_t(t_j, (cy_1, \ldots, cy_t))$. The same is true for other kinds of inequalities. We get $cy_t > M_t(t_j, (cy_1, \ldots, cy_t))$ for $c < 0$. I.e., the absolute value of statistics $A^t_{t_j+k}(M_t(t_j))$ for $k = t - t_j, \ldots, 1$ is independent of whether we use the sample $y_1, \ldots, y_n$ or $cy_1 + s, \ldots, cy_n + s$ for $c, s \in \mathbb{R}$ and $c \neq 0$. The inequality in Step 12 does not depend on the shift or scale of the sample under these conditions. The case $c = 0$ is simple. □

We use the statistics $A^{t+k}_t(M_{t+k}(t_j))$ because of their simplicity and the possibility of recursive calculations, see Remark 2.2.1. The computation of all relevant A^k_t for each t in our algorithm has a computational complexity at worst n.

We know that sample median converges to its theoretical counterpart for i.i.d. random variables, see e.g., (Pollard, 2012). I.e., if we do not compute the statistics $A^t_{1+k}(M_t(1))$ for $k = t - 1, \ldots, 1$, do not look for the change point in Step 8 of Algorithm 2.2.1 and there are no real change points in the data, then $\hat{a}_t \xrightarrow{a.s.} a$, since ε_t has its median equal to 0. From there follows also the consistency.

Proposition 2.2.4. *Consider an infinite series $y_1, y_2, \ldots$ without level shifts following Model* (2.19) *i.e.,*

$$y_t = a + \varepsilon_t$$

and suppose further that ε_t are non-degenerate random variables. Then Algorithm 2.2.1 identifies infinitely many points as level shifts almost surely.

Proof. Due to the assumption that ε_t has the theoretical median equal to 0 and that it is not degenerated random variable, there are $\delta > 0$ and $c_1 > 0$ such that $\mathrm{P}(y_t > a + \delta) > c_1$ or $\mathrm{P}(y_t < a - \delta) > c_1$. We suppose without loss of generality that the $\mathrm{P}(y_t > a + \delta) > c_1$. If only the case $\mathrm{P}(y_t < a - \delta) > c_1$ is valid then the proof proceeds in the same fashion with some small exceptions.

Divide the infinite series $y_1, y_2, \ldots$ to parts of a length $k \in \mathbb{N}$ such that the statistics $A^{(i+1)k}_{ik+1}(a + \delta)$ for $i = 0, 1, \ldots$ are greater than b in the case that all

$I_{ik+j}(a+\delta) = 1$ for $i = 0, 1, \ldots$ and $j = 1, \ldots, k$. We consider these subsequences as distinct.

Compute statistics $A_{ik+1}^{(i+1)k}(a+\delta)$ for $i = 0, 1, \ldots$ These statistics are independent, because the observations do not overlap among the subsequences. Furthermore, the number of these statistics is infinite. There exists $c_2 \in (0, 1)$ such that for any $i \in \mathbb{N}$ holds

$$\mathrm{P}\left(A_{ik+1}^{(i+1)k}(a+\delta) > b\right) > c_2 > 0.$$

This with the help of the Borel-Cantelli lemma implies that there are infinitely many i for which

$$A_{ik+1}^{(i+1)k}(a+\delta) > b.$$

We know that $M_t(1) \xrightarrow{a.s.} a$ which yields that there is $k_0 \in \mathbb{N}$ such that for all $k \in \mathbb{N}$, $k \geq k_0$ and $i \in \mathbb{N}$ holds $M_{(i+1)k}(1) \leq a + \delta$ almost surely. From that follows that if $A_{ik+1}^{(i+1)k}(a+\delta) > b$ then $A_{ik+1}^{(i+1)k}(M_{(i+1)k}(1)) > b$ for $k \geq k_0$.

It means that infinitely many potential level shifts appear but only for the split time series.

We proceed by a mathematical induction in the rest of the proof. We know from the previous that for a time series which fulfills our assumptions we always find some level shift no later than at time $ik+1$ for which $A_{ik+1}^{(i+1)k}(M_{(i+1)k}(1)) > b$ is fulfilled. Let t_1 denote the time, where the first level shift appears (it holds $t_1 \leq ik+1$). Let us suppose now, that we have found the level shift in t_j and show that there is one in t_{j+1}. But the time series $y_{t_j+1}, y_{t_j+2}, \ldots$ fulfills also our assumptions, which means that there is a level shift in t_{j+1}. □

Remark 2.2.5. *Algorithm 2.2.1 converges only if $b = \infty$ according to Proposition 2.2.4, i.e., we do not check the statistics $A_{1+k}^t(M_t(1))$ for $k = t-1, \ldots, 1$. Intuitively, it is clear that our algorithm converges "better" for higher b. On the other hand higher b means lower sensitivity to level shifts.*

Further, we want to study the robustness of Algorithm 2.2.1 in the sense of Definition 1.1.7. The number of observations $m(t, \boldsymbol{y})$ from which the final estimate $\hat{a}_t$ is computed may vary. Due to the properties of median and employing Definition 1.1.6 we get $\varepsilon^*_{t_j - t_{j-1}}(M_{t_j-1}(t_{j-1}), (y_{t_{j-1}}, \ldots, y_{t_j-1})^\top) = \frac{1}{t_j - t_{j-1}} \left\lfloor \frac{t_j - t_{j-1} - 1}{2} \right\rfloor$. We employ the notation $M_t(1, (y_1, \ldots, y_t))$ instead of the usual $M_t(1)$ to stress the role of the observations y_t. We define

$$k_{b,W} = \min_{k=t,\ldots,1; t \geq W; t, W \in \mathbb{N}} \left\{ k \,\middle|\, \left|A_k^t(M_t(1, (y_1, \ldots, y_t)))\right| > b; y_1, \ldots, y_t \in \mathbb{R}^t \right\} - 1, \tag{2.22}$$

where only one observation can always attain the value of median. I.e., there is maximally one i among $y_1, \ldots, y_t$ and $t \in \mathbb{N}$ for which $y_i = M_t(1, (y_1, \ldots, y_t))$. The last condition prevents observations from attaining median values and so reduce $k_{b,W}$. This is handy in Proposition 2.2.7 for continuous random variables. $k_{b,W}$ expresses, under these conditions, the least possible number of observations from which final estimates $\hat{a}_t$ are computed. This follows from Step 10 of Algorithm 2.2.1.

Example 2.2.6. *We focus our attention on $k_{b,W}$ in this example. We consider without loss of generality the case when $A_k^t(M_t(1))$ exceeds b, because too many* *observations are higher than median $M_t(1)$. We also suppose for simplicity that no observation is equal to $M_t(1)$ (i.e., t is even). Let us fix $t \in \mathbb{N}$. We know that $\frac{t}{2}$ of observations has to be above $M_t(1)$ and $\frac{t}{2}$ below $M_t(1)$. We look for the least number of observations s_τ (we simplify the notation by neglecting $t, b, (y_1, \ldots, y_t)$ on which s_τ also depends) which has to be above the median $M_t(1)$ in order to fulfill $A^t_{t-\tau+1}(M_t(1)) \geq b$ for $\tau \leq t$ and $\tau \in \mathbb{N}$. We get from* (2.21)

$$A^t_{t-\tau+1}(M_t(1)) = \frac{2S^t_{t-\tau+1}(M_t(1)) - \tau}{\sqrt{\tau}} \geq b.$$

It yields

$$S^t_{t-\tau+1}(M_t(1)) \geq \frac{b\sqrt{\tau} + \tau}{2}. \tag{2.23}$$

According to (2.20)*, $S^t_{t-\tau+1}(M_t(1))$ is the number of observations which are above $M_t(1)$. This gives $s_\tau = \frac{b\sqrt{\tau}+\tau}{2}$. If we need a number from $\mathbb{N}$, we have to take $\left\lceil \frac{b\sqrt{\tau}+\tau}{2} \right\rceil$. We define*

$$\bar{k}_{b,t} = \min_{k=t,\ldots,1} \left\{ k \middle| \left| A^t_k(M_t(1)) \right| > b \right\} - 1.$$

We know from the previous considerations that $(y_{\bar{k}_{b,t}+1}, \ldots, y_t)^\top$ has to contain at least $s_{t-\bar{k}_{b,t}+1}$ observations, which are above median $M_t(1)$. It means that among $y_1, \ldots, y_{\bar{k}_{b,t}}$ there has to be at least

$$\frac{t}{2} - (t - \bar{k}_{b,t} + 1 - s_{t-\bar{k}_{b,t}+1}) = \frac{\bar{k}_{b,t} - 1 + b\sqrt{t - \bar{k}_{b,t} + 1}}{2} \tag{2.24}$$

observations which are below $M_t(1)$ to gain $\frac{t}{2}$ of observations below $M_t(1)$. From there and from the fact that the total number of observations among $y_1, \ldots, y_{\bar{k}_{b,t}}$ has to be higher than the number of observations below $M_t(1)$, we get $\bar{k}_{b,t} \geq \frac{\bar{k}_{b,t}-1+b\sqrt{t-\bar{k}_{b,t}+1}}{2}$. This gives

$$\bar{k}^2_{b,t} + \bar{k}_{b,t}(2 + b^2) - b^2(t + 1) + 1 \geq 0. \tag{2.25}$$

The roots of (2.25) *are $\frac{-(2+b^2)\pm b\sqrt{b^2+4t+8}}{2}$. For $\bar{k}_{b,t} > 0$, we know*

$$\bar{k}_{b,t} \geq \frac{-(2 + b^2) + b\sqrt{b^2 + 4t + 8}}{2}. \tag{2.26}$$

The second solution is always negative.

To be precise, we should employ in (2.24) $\left\lceil s_{t-\bar{k}_{b,t}+1} \right\rceil$ *to get*

$$\bar{k}_{b,t} \geq \frac{t}{2} - \left(t - \bar{k}_{b,t} + 1 - \left\lceil s_{t-\bar{k}_{b,t}+1} \right\rceil \right),$$

for it expresses the minimal naturel number of observations which have to be higher than $M_t(1)$. It has to be also fulfilled that $\frac{-(2+b^2)+b\sqrt{b^2+4t+8}}{2} \leq t$, for $\bar{k}_{b,t}$ has to be less or equal to t, otherwise there is no solution. Nevertheless, (2.26) *yields that the least possible value for $\bar{k}_{b,t}$ grows with growing t. Therefore, it seems reasonable to choose t as low as possible to gain the possible lowest $k_{b,t}$, namely $t = W$.* △

Proposition 2.2.7. *Consider an infinite series $y_1, y_2, \dots$ without level shifts following Model* (2.19) *i.e.,*

$$y_t = a + \varepsilon_t$$

and suppose further that ε_t are non-degenerated continuous random variables. Let $T_n : \mathbb{R}^n \to \mathbb{R}^n$ be an estimator defined by Algorithm 2.2.1 with inputs $\boldsymbol{y} = (y_1, \dots, y_n)^\top$, $W \in \mathbb{N}$, $W \leq n$, $b > 0$ and $T_n \to \bar{T}$ for $n \to \infty$. Then

$$\varepsilon(\bar{T}, \boldsymbol{y}) = \frac{1}{k_{b,W}} \left\lfloor \frac{k_{b,W} - 1}{2} \right\rfloor, \tag{2.27}$$

where $\varepsilon(\bar{T}, \boldsymbol{y})$ is defined in Definition 1.1.7, $k_{b,W}$ by 2.22 and $\lfloor \cdot \rfloor$ is the floor function.

Proof. We know that $\varepsilon^*_{m(t,\boldsymbol{y})}(\bar{T}_t, \boldsymbol{z}_t) = \frac{1}{m(t,\boldsymbol{y})} \left\lfloor \frac{m(t,\boldsymbol{y})-1}{2} \right\rfloor$, where $m(t, \boldsymbol{y}) \geq k_{b,W}$. For $\frac{1}{m} \left\lfloor \frac{m-1}{2} \right\rfloor$ is a nondecreasing sequence with respect to m, we have proved that $\varepsilon(\bar{T}, \boldsymbol{y}) \geq \frac{1}{k_{b,W}} \left\lfloor \frac{k_{b,W}-1}{2} \right\rfloor$. We show to prove the equality in (2.27) that $k_{b,W}$ is the minimal $m(t, \boldsymbol{y})$ almost surely. We have shown in Proposition 2.2.4 that there are infinitely many change points found. Consider two change points one in time t_i and the second in time t_j such that $t_i \neq t_j$. The observations in sequences $y_{t_j}, \dots, y_{t_{j+1}-1}$ and $y_{t_i}, \dots, y_{t_{i+1}-1}$ are independent. We note that t_i and t_j are not independent; nevertheless, $t_{j+1} - t_j$ and $t_{i+1} - t_i$ are. We denote the length between two change points $k_i = t_{i+1} - t_i$. For the assumption that y_τ is continuous for any $\tau \in \mathbb{N}$ maximally one $i = 1, \dots, t$ can fulfill $y_i = M_t(1)$. Therefore, the assumptions of (2.22) are satisfied. It yields there is nonzero probability $\mathrm{P}\,(\text{the first change point is in } k_{b,W}) = c > 0$. We know k_i for $i = 1, 2, \dots$ are independent and $\sum_{i=1}^{\infty} \mathrm{P}(k_i = k_{b,W}) = \sum_{i=1}^{\infty} c = \infty$. We can conclude $\mathrm{P}(k_i = k_{b,W} \text{ i.o.}) = 1$ with the help of the Borel-Cantelli lemma. Thus, we have proved that there are infinitely many t for which $m(t, \boldsymbol{y}) = k_{b,W}$ almost surely. □

The proposition cannot be proved for any random variable ε_t for the possibility of attaining median more times than once by some observations. If we loosen the assumption for $k_{b,W}$ (2.22) about attaining the value of median then we can prove in Proposition 2.2.7 only the inequality $\varepsilon(\bar{T}, \boldsymbol{y}) \geq \frac{1}{k_{b,W}} \left\lfloor \frac{k_{b,W}-1}{2} \right\rfloor$ and the value of $k_{b,W}$ would be lower.

We can want to omit the condition about attaining median in 2.22 and define.

$$\bar{k}_{b,W} = \min_{k=t,\dots,1; t \geq W; t,W \in \mathbb{N}} \left\{ k \,\middle|\, \left| A_k^t(M_t(1, (y_1, \dots, y_t))) \right| > b; y_1, \dots, y_t \in \mathbb{R}^t \right\} - 1. \tag{2.28}$$

We employ Definition 1.1.8 in the following proposition and therefore we do not have to make any assumptions connecting to the distribution of ε_t (we still suppose that median of ε_t is equal to zero).

Proposition 2.2.8. *Consider an infinite series $y_1, y_2, \dots$ without level shifts following Model* (2.19) *i.e.,*

$$y_t = a + \varepsilon_t.$$

Let $T_n : \mathbb{R}^n \to \mathbb{R}^n$ be an estimator defined by Algorithm 2.2.1 with inputs $\boldsymbol{y} = (y_1, \dots, y_n)^\top$, $W \in \mathbb{N}$, $W \leq n$, $b > 0$ and $T_n \to \bar{T}$ for $n \to \infty$. Then

$$\varepsilon(\bar{T}, \boldsymbol{y}) = \frac{1}{\bar{k}_{b,W}} \left\lfloor \frac{\bar{k}_{b,W} - 1}{2} \right\rfloor, \tag{2.29}$$

where $\varepsilon(\bar{T}, \boldsymbol{y})$ is defined in Definition 1.1.8, $k_{b,W}$ by 2.22 and $\lfloor \cdot \rfloor$ is the floor function.

Proof. The first inequality can be proved in a similar way as in Proposition 2.2.7. I.e., it came from Equation (2.28) and the fact that the function $\frac{1}{m}\lfloor \frac{m-1}{2} \rfloor$ is a nondecreasing sequence with respect to m.

The second inequality came from the fact that the algorithm itself enables $m(t, \boldsymbol{y})$ to be equal to $\bar{k}_{b,W}$ and therefore there is $\boldsymbol{z} \in S(t, \boldsymbol{y})$ (see Definition 1.1.8) such that the number of observations in $\boldsymbol{z}$ is equal to $\bar{k}_{b,W}$. □

We compute now $\pi_t(T_n)$ from Definition 1.1.10, where T_n is defined according to Algorithm 2.2.1 and the rest according to the definition. We get

$$\pi_t(T_n(\bar{\boldsymbol{y}})) = \sum_{i=0}^{\lfloor \frac{m(t,\boldsymbol{y})-1}{2} \rfloor} p^i (1-p)^{m(t,\bar{\boldsymbol{y}})-i}.$$

I.e., we are looking for the probability that less than $\left\lfloor \frac{m(t,\boldsymbol{y})-1}{2} \right\rfloor$ observations are violated, where $\left\lfloor \frac{m(t,\boldsymbol{y})-1}{2} \right\rfloor + 1$ expresses the number of contaminated observations because of which $\hat{a}_t$ can reach infinity.

We want to discuss a time complexity, see Definition 1.1.13, of Algorithm 2.2.1 now. We consider the worst case meaning that no level shifts are found. I.e., we have to compute the median from all observations up to the time t in each step. We can exploit the information about ordering from the previous step and therefore we get the time complexity $\mathcal{O}(t)$ in each step. In the last step $n = t$; therefore, we replace t by n and get the time complexity of the algorithm $\mathcal{O}(n^2)$.

The time complexity depends on b in reality (also on the number of real level shifts in the series); if b is small then we have to order only a few observations and so the complexity is close to $\mathcal{O}(n)$.

2.2.2 Sign test: General case

We return to the general model (2.1) and sketch a rough idea of an algorithm which can be employed in such a general case. Then, we focus only on a special case of the linear trend.

We mean by the series of **pre-estimates** an auxiliary series of preliminary estimates of parameters obtained in each time t. I.e., it includes a vector of d components in each time t. These are not the final estimates of parameters but they enable us to construct such final estimates.

Consider $\boldsymbol{r} = (r_0, \ldots, r_{d-1})^\top \in \mathbb{R}^d$, we introduce a notation $\boldsymbol{r}_{\cdot i} = (r_0, \ldots, r_{i-1}, r_{i+1}, \ldots, r_{d-1})^\top$ for $i = 0, \ldots, d-1$. Let us suppose that we have robust parameter estimates $\hat{\boldsymbol{a}}_{t-1}((y_1, \ldots, y_n), \Theta_{t-1})$ from the previous step, where Θ_{t-1} is a vector of parameters on which the estimate depends. We neglect $((y_1, \ldots, y_n), \Theta_t)$ in the following text and simply write $\hat{\boldsymbol{a}}_t$ for the robust parameter estimate at the time t. We further denote by $\bar{\boldsymbol{a}}_t(y_t, \boldsymbol{z}_t, \hat{\boldsymbol{a}}_{t-1})$ the vector of pre-estimates at the time t. The i-th component $\bar{a}_{i,t}(y_t, \boldsymbol{z}, \hat{\boldsymbol{a}}_{t-1})$ of $\bar{\boldsymbol{a}}_t(y_t, \boldsymbol{z}_t, \hat{\boldsymbol{a}}_{t-1})$ for $i = 0, \ldots, d-1$ is a solution of the equation

$$y_t = z_{i,t}\bar{a}_{i,t}(y_t, \boldsymbol{z}_t, \hat{\boldsymbol{a}}_{t-1}) + \boldsymbol{z}_{\cdot i,t}^\top \hat{\boldsymbol{a}}_{\cdot i,t-1}. \tag{2.30}$$

We gain all components of the vector of the pre-estimates at time t in this way.

Let the last found change point is at time t_j. Then $\hat{a}_{i,t} = M_t(t_j, (\bar{a}_{i,t_j}, \dots, \bar{a}_{i,t}))$ for $i = 0, \dots, d-1$, where we neglect $(y_t, \boldsymbol{z}_t, \hat{\boldsymbol{a}}_{t-1})$ by $\bar{a}_{i,t}(y_t, \boldsymbol{z}_t, \hat{\boldsymbol{a}}_{t-1})$. I.e., we gain $\hat{\boldsymbol{a}}_t$ as the multidimensional median from the pre-estimates $\bar{\boldsymbol{a}}_{t_j}, \dots, \bar{\boldsymbol{a}}_t$.

We proceed in the same fashion as in 2.2.1 to find the next change point. I.e., we put $\hat{y}_{\tau,t}(\hat{\boldsymbol{a}}_t, \boldsymbol{z}_t) = \boldsymbol{z}_t^\top \hat{\boldsymbol{a}}_t$ for $\tau = t_j, \dots, t$. If too many consecutive estimates $\hat{y}_{\tau,t}$ for $\tau = 1, \dots, t$ lie above or under y_τ, we proclaim the first point τ, where such behavior appears, to be a new change point.

The details of the algorithm are described for a special case of the linear trend.

2.2.3 Sign test: Linear trend

Consider the model with the linear trend

$$y_t = a_{0,t} + a_{1,t}t + \varepsilon_t, \qquad \text{for } t = 1, \dots, n. \tag{2.31}$$

We extend the notation and assumptions from part 2.2.1. Let $\boldsymbol{c}_{t_j}, \dots, \boldsymbol{c}_t$ be a series of vectors $\boldsymbol{c}_i \in \mathbb{R}^2$ for $i = t_j, \dots, t$. We denote by $\boldsymbol{M}_t(t_j, (\boldsymbol{c}_{t_j}, \dots, \boldsymbol{c}_t))$ the componentwise median from $\boldsymbol{c}_{t_j}, \dots, \boldsymbol{c}_t$.

We utilize a fixed $W \in \mathbb{N}$ denoting the number of observations employed after each change point to compute the initial estimates as in the case of constant trend 2.2.1. We also employ a constant $b > 0$ relevant for the test statistics. The interval $(-b, b)$ and W are found by the simulation experiments presented in 2.3.2.

Consider $\hat{\boldsymbol{a}}_t((y_1, \dots, y_t), b, W)$ an estimate of $\boldsymbol{a}_t = (a_{0,t}, a_{1,t})^\top$ valid at time t found by means of our recursive Algorithm 2.2.2. We usually simplify the notation and write only $\hat{\boldsymbol{a}}_t = (\hat{a}_{0,t}, \hat{a}_{1,t})^\top$.

We need initial estimates of $a_{0,t}$ and $a_{1,t}$ from (2.31) in the beginning and after each found change point. We denote them $\tilde{a}_{i,t_j}(y_{t_j}, \dots, y_{t_j+k})$ for $i = 0, 1$ and $k \in \mathbb{N}$. If $k = W$ we employ simpler notation $\tilde{a}_{i,t_j} = \tilde{a}_{i,t_j}(y_{t_j}, \dots, y_{t_j+W})$ for $i = 0, 1$ or for vectors $\tilde{\boldsymbol{a}}_{t_j} = (\tilde{a}_{0,t_j}, \tilde{a}_{1,t_j})^\top$. These estimates are not obtained recursively, like the estimates $\hat{a}_{0,t}$ and $\hat{a}_{1,t}$, but from the first W observations after a change point. To find $\tilde{\boldsymbol{a}}_{t_j}$ we employ

$$\tilde{\boldsymbol{a}}_{t_j} \in \operatorname*{argmin}_{\boldsymbol{c} \in \mathbb{R}^2} \sum_{t=t_j}^{t_j+W} |y_t - c_0 - c_1 t|. \tag{2.32}$$

We want to stress by the symbol $\in$ that $\tilde{\boldsymbol{a}}_{t_j}$ belongs among the solutions of the expression on the right side of (2.32).

We deal only with two parameters; so the pre-estimates form two series. The members of these series of pre-estimates for $\boldsymbol{a}_t$ are denoted by $\bar{\boldsymbol{a}}_t(\boldsymbol{c})$, where $\boldsymbol{c} = (c_0, c_1)^\top$ for some $c_0, c_1 \in \mathbb{R}$. We denote by $\bar{a}_{0,t}(\boldsymbol{c})$ the solution of

$$y_t = \bar{a}_{0,t}(\boldsymbol{c}) + c_1 t \tag{2.33}$$

and as $\bar{a}_{1,t}(\boldsymbol{c})$ the solution of

$$y_t = c_0 + \bar{a}_{1,t}(\boldsymbol{c}) t. \tag{2.34}$$

We define statistics $S_t^{t+k}(c_0 + c_1 t)$ and $A_t^{t+k}(c_0 + c_1 t)$ for suitable $k \in \mathbb{N}$ in a similar way as (2.20) and (2.21), respectively:

$$S_t^{t+k}(c_0 + c_1\tau) = \sum_{\tau=t}^{t+k} I_\tau(c_0 + c_1\tau)$$

and

$$A_t^{t+k}(c_0 + c_1\tau) = \frac{2S_t^{t+k}(c_0 + c_1\tau) - k - 1}{\sqrt{k+1}}. \tag{2.35}$$

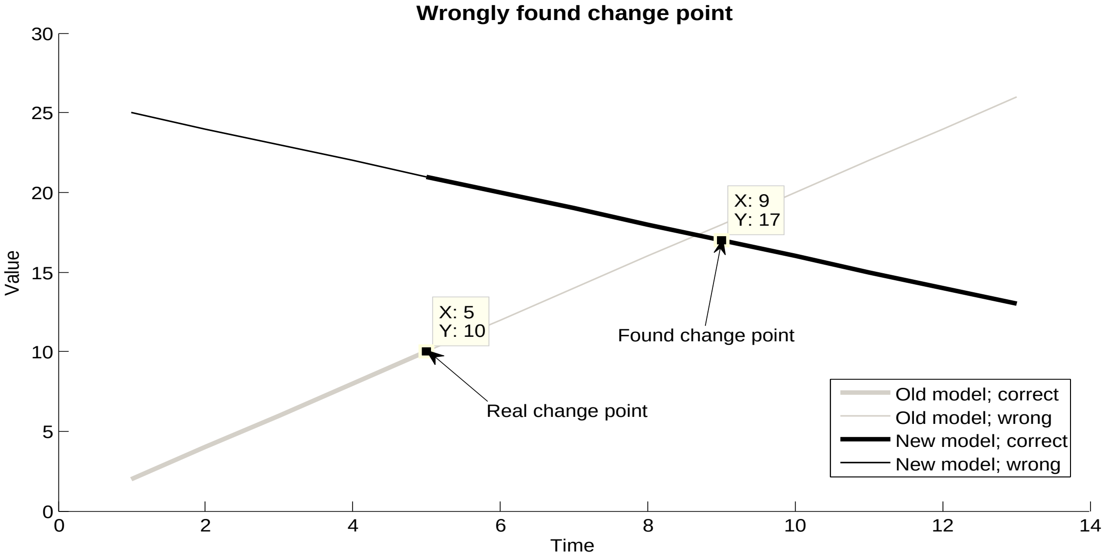


Figure 2.1. (linear trend): Possibly wrongly identified change point in steps 1 - 14 of Algorithm 2.2.2.

We consider now the real change point to appear at time τ. Let us deal with a situation when the found change point t_j from Steps 1 - 14 of Algorithm 2.2.2 is misspecified. The situation is more problematic than in Lemma 2.2.2, for the case shown in Figure 2.1.

Example 2.2.9. *We describe the situation depicted in Figure 2.1 in this example. We suppose that $\varepsilon_t = 0$ for $t = 1, \ldots, n$. y_t follows the model*

$$y_t = a_0^o + a_1^o t, \qquad \text{for } t = 1, \ldots, \tau - 1$$

up to the time $\tau - 1$.

It follows

$$y_t = a_0^n + a_1^n t, \qquad \text{for } t = \tau, \ldots, n$$

after the time $\tau - 1$.

Let

$$a_0^n + a_1^n r = a_0^o + a_1^o r,$$

for some $r \in \mathbb{R}$ and $r > \tau$. Let further

$$b > \left| A_\tau^{\lceil r \rceil}(a_0^o + a_1^o\theta) \right|,$$

where $\lceil \cdot \rceil$ denotes the ceiling function.

Steps 1-14 identify the change point t_j such that $r \in [t_j - 1, t_j)$. △

Nevertheless, the situation from Lemma 2.2.2 can also appear. We skip the simple proof.

Lemma 2.2.10. *Consider the case with a real change point at time* $\tau \in \mathbb{N}$, $\tau > t_j$ *and the Model* (2.19) *changing from*

$$y_t = a_0^o + a_1^o t, \qquad \text{for } t = 1, \dots, \tau - 1$$

to

$$y_t = a_0^n + a_1^n t, \qquad \text{for } t = \tau, \dots, n.$$

In the case that $y_{\tau-1} > \hat{a}_{0,t} + \hat{a}_{1,t}(\tau - 1)$ *and* $A_\tau^t(\hat{a}_{0,t} + \hat{a}_{1,t} i) > b$ *for some* $t \in \mathbb{N}$ *and* $t > \tau$ *then also* $A_{\tau-1}^t(\hat{a}_{0,t} + \hat{a}_{1,t} i) > b$. *This arises especially in the case of* $a_0^n + a_1^n \tau > a_0^o + a_1^o \tau$.

In the opposite case, when $y_{\tau-1} < \hat{a}_{0,t} + \hat{a}_{1,t}(\tau - 1)$ *and* $A_\tau^t(\hat{a}_{0,t} + \hat{a}_{1,t} i) < -b$ *then* $A_{\tau-1}^t(\hat{a}_{0,t} + \hat{a}_{1,t} i) < -b$. *This arises especially in the case of* $a_0^n + a_1^n \tau < a_0^o + a_1^o \tau$.

We plug Steps 17-24 into Algorithm 2.2.2 to deal with these obstacles described in Example 2.2.9 and Lemma 2.2.10.

Algorithm 2.2.2 Algorithm for a linear trend based on the sign test

```
Input: observations $(y_1, \dots, y_n)^\top$, $W \in \mathbb{N}$, $b > 0$
 1: put $t \leftarrow 1$, $j \leftarrow 1$, $t_1 \leftarrow 1$
 2: while $t \leq n$ do          ▷ this has to be checked throughout the While loop
 3:     compute $\tilde{\boldsymbol{a}}_{t_j}$
 4:     for $\tau = t_j, \dots, t_j + W$ do
 5:         $\hat{\boldsymbol{a}}_\tau \leftarrow \tilde{\boldsymbol{a}}_{t_j}$
 6:         compute $\bar{\boldsymbol{a}}_\tau(\hat{\boldsymbol{a}}_\tau)$
 7:     end for
 8:     $t \leftarrow t_j + W - 1$
 9:     repeat
10:         $t \leftarrow t + 1$
11:         compute $\bar{\boldsymbol{a}}_t(\hat{\boldsymbol{a}}_{t-1})$
12:         $\hat{\boldsymbol{a}}_t \leftarrow \boldsymbol{M}_t\left(t_j, \left(\bar{\boldsymbol{a}}_{t_j}(\hat{\boldsymbol{a}}_{t_j}), \bar{\boldsymbol{a}}_{t_j+1}(\hat{\boldsymbol{a}}_{t_j}), \dots, \bar{\boldsymbol{a}}_{t-1}(\hat{\boldsymbol{a}}_{t-2}), \bar{\boldsymbol{a}}_t(\hat{\boldsymbol{a}}_{t-1})\right)\right)$
13:         compute $A_{t_j+k}^t(\hat{a}_{0,t} + \hat{a}_{1,t} t)$ for $k = t - t_j, \dots, 1$
14:     until $A_{t_j+k}^t(\hat{a}_{0,t} + \hat{a}_{1,t} \tau) \in (-b, b)$ for $k = t - t_j, \dots, 1$
```

```
    $j \leftarrow j + 1$
    $t_j \leftarrow t_{j-1} + \min\left\{ k \middle| \left| A^t_{t_{j-1}+k}(\hat{a}_{0,t} + \hat{a}_{1,t}\tau) \right| > b; k = t - t_{j-1}, \dots, 1 \right\}$
    compute $\tilde{\boldsymbol{a}}_{t_j}$
    while $|\tilde{a}_{0,t_j} + \tilde{a}_{1,t_j}(t_j - 1) - y_{t_j-1}| < |\hat{a}_{0,t_j-1} + \hat{a}_{1,t_j-1}(t_j - 1) - y_{t_j-1}|$ and $t_j > t_{j-1}$ do
        $t_j \leftarrow t_j - 1$
    end while
    while $|\tilde{a}_{0,t_j} + \tilde{a}_{1,t_j}(t_j + 1) - y_{t_j+1}| > |\hat{a}_{0,t_j-1} + \hat{a}_{1,t_j-1}(t_j + 1) - y_{t_j+1}|$ and $t_j < t$ do
        $t_j \leftarrow t_j + 1$
        compute $\tilde{\boldsymbol{a}}_{t_j}$
    end while
    if $t_j - t_{j-1} < W$ then
        compute $\tilde{\boldsymbol{a}}_{t_{j-1}}(y_{t_{j-1}}, \dots, y_{t_j-1})$
        for $t = t_{j-1}, \dots, t_j - 1$ do
            $\hat{y}_t \leftarrow \tilde{a}_{0,t_{j-1}}(y_{t_{j-1}}, \dots, y_{t_j-1}) + \tilde{a}_{1,t_{j-1}}(y_{t_{j-1}}, \dots, y_{t_j-1})t$
        end for
    else
        $y_t \leftarrow \hat{a}_{0,t_j-1} + \hat{a}_{1,t_j-1}t$ for $t = t_{j-1}, \dots, t_j - 1$
    end if
end while
```

The following text describes possible adjustments of Algorithm 2.2.2. We can replace by some other robust estimate of location the componentwise median serving as an estimate of $\boldsymbol{a}_t$, e.g., by M-estimators (Huber, 1964), τ-estimators (Lopuhaä, 1991) or spatial median (Kemperman, 1987). Componentwise median is a special case of spatial median in l_1 norm. These methods are studied in Chapter 3. We choose componentwise median especially because of its simplicity and robust properties (breakdown point equal to $\frac{1}{2}$).

We need a robust regression estimator to compute $\tilde{\boldsymbol{a}}_{t_j}$. It can be performed in many different ways. The alternatives to the method mentioned in the beginning of this section were already discussed in Chapter 1, e.g., the weighted least squares (Víšek, 2000) or Theil-Sen algorithm (Theil, 1950). Although our method for choosing $\tilde{\boldsymbol{a}}_{t_j}$ is really simple, it does not achieve the highest possible breakdown point. We should utilize the variation of the Theil-Sen algorithm described in (Siegel, 1982) to reach it.

We consider two options how to deal with the statistics A_t^{t+k} where $t, k \in \mathbb{N}$.

(a) We can apply statistics (2.21) on the series of pre-estimates $\bar{a}_{0,t}(\hat{\boldsymbol{a}}_{t-1})$ and $\bar{a}_{1,t}(\hat{\boldsymbol{a}}_{t-1})$. But it is clear from Equation (2.36) and (2.37) and Example 2.2.11 that the series of pre-estimates can be biased. Moreover, the median computed from them can also be biased. I.e., we would not be able to reveal the wrong estimation of parameters $\boldsymbol{a}_t$ after employing these statistics.

(b) The second option is to employ statistics (2.35). It has the disadvantage in comparison to Option a that is described in Example 2.2.9. I.e., it does not have to show the proper time of the change point. It also does not reveal which component of $\boldsymbol{a}$ has changed, which can be useful. However, we compare our estimated values $\hat{a}_{0,t} + \hat{a}_{1,t}t$ directly to observations y_t and; therefore, we reveal any bias much faster. So, we prefer this option.

It is possible not to recompute the values of $\hat{a}_{0,t_j}$ and $\hat{a}_{1,t_j}$ in Steps 21 and 18 if we keep the values for $\tau > t_{j-1}$ in the memory. It is also possible to employ instead the estimates of $\hat{a}_{0,t}$ and $\hat{a}_{1,t}$ for such t for which these values were computed the last time, but such a simplistic approach could give worse results.

There is another option for Steps 18 - 20. We can, instead of comparing the values $|\tilde{a}_{0,t_j}+\tilde{a}_{1,t_j}(t_j-1)-y_t|$ and $|\hat{a}_{0,t_j}+\hat{a}_{1,t_j}(t_j-1)-y_t|$, recompute the statistics 2.35 in the following way. We define for $\boldsymbol{n} = (n_0, n_1)^\top \in \mathbb{R}^2$, $\boldsymbol{o} = (o_0, o_1)^\top \in \mathbb{R}^2$ and $o \in \mathbb{R}$

$$I_{t,\boldsymbol{n}}(o) = \begin{cases} 1 & \text{if } (y_t > o \text{ and } n_0 + n_1 t > o) \text{ or } (y_t < o \text{ and } n_0 + n_1 t < o), \\ \frac{1}{2} & \text{if } y_t = o, \\ 0 & \text{if } (y_t > o \text{ and } n_0 + n_1 t \le o) \text{ or } (y_t < o \text{ and } n_0 + n_1 t \ge o). \end{cases}$$

Further, similarly to 2.35

$$S_{t,\boldsymbol{n}}^{t+k}(o_0 + o_1\tau) = \sum_{\tau=t}^{t+k} I_{\tau,\boldsymbol{n}}(o_0 + o_1\tau)$$

and

$$A_{t,\boldsymbol{n}}^{t+k}(o_0 + o_1\tau) = \frac{2S_{t,\boldsymbol{n}}^{t+k}(o_0 + o_1\tau) - k - 1}{\sqrt{k+1}}.$$

Algorithm 2.2.3 Alternative to Steps 18 - 20 of Algorithm 2.2.2

```
18: compute $A^{t}_{t_{j-1}+k,\tilde{\boldsymbol{a}}_{t_j}}(\hat{a}_{0,t} + \hat{a}_{1,t}t)$ for $k = t - t_{j-1}, \ldots, 1$
19: $A^{t}_{t,\tilde{\boldsymbol{a}}_{t_j}} \leftarrow b + 1$
20: $t_j \leftarrow t_{j-1} + \min\left\{k | A^{t}_{t_{j-1}+k,\tilde{\boldsymbol{a}}_{t_j}}(\hat{a}_{0,t} + \hat{a}_{1,t}\tau) > b; k = t - t_{j-1}, \ldots, 1\right\}$
```

This alternative was implemented in the code for Simulation study 2.3.

We deal with the convergence of Algorithm 2.2.2 and its other properties in the following part. A similar idea to the implementation of the series of pre-estimates was presented in (Holt, 2004). They simply put $\bar{a}_{1,t} = y_t - y_{t-1}$ in that paper. Such an approach has the disadvantage that it is not robust in the sense that an outlier influences two members of the series of pre-estimates. This can be solved by replacing y_{t-1} by its robust estimate $\hat{y}_{t-1} = \hat{a}_{0,t-1} + \hat{a}_{1,t-1}(t-1)$ (compare (Gelper et al., 2010)). However, it is also inappropriate to put $\bar{a}_{1,t}(\hat{\boldsymbol{a}}_{t-1}) = y_t - \hat{a}_{0,t-1} - \hat{a}_{1,t-1}(t-1)$, since the residuals are summed up in this case while they are divided by t in Algorithm 2.2.2. Denote $\delta_{0,t-1} = \hat{a}_{0,t-1} - a_{0,t-1}$ and $\delta_{1,t-1} = \hat{a}_{1,t-1} - a_{1,t-1}$. Suppose further that there is no change point for several observations around the time t so that the parameters are constant. Then

$$\begin{aligned} y_t - \hat{a}_{0,t-1} - \hat{a}_{1,t-1}(t-1) &= a_{1,t} + (a_{0,t} - \hat{a}_{0,t-1}) + (a_{1,t} - \hat{a}_{1,t-1})(t-1) + \varepsilon_t \\ &= a_{1,t} + \delta_{0,t-1} + \delta_{1,t-1}(t-1) + \varepsilon_t \end{aligned}$$

in comparison to

$$\bar{a}_{1,t}(\hat{\boldsymbol{a}}_{t-1}) = \frac{y_t - \hat{a}_{0,t-1}}{t} = \frac{\delta_{0,t-1} + a_{1,t}t + \varepsilon_t}{t} = a_{1,t} + \frac{\delta_{0,t-1} + \varepsilon_t}{t}. \tag{2.36}$$

Nevertheless, there still remains a pitfall

$$\bar{a}_{0,t}(\hat{\boldsymbol{a}}_{t-1}) = y_t - \hat{a}_{1,t-1}t = a_{0,t} + \varepsilon_t + \delta_{1,t-1}t. \tag{2.37}$$

Equation (2.37) shows that errors from previous steps do not have to shrink. This presents a big disadvantage of Algorithm 2.2.2, for it does not have to converge as it is demonstrated in Example 2.2.11.

We know the sample median of i.i.d. random variables converges almost surely to its theoretical counterpart, see e.g., (Pollard, 2012). The case is studied when the random sample median converges to its counterpart and the random variables are not identically distributed in (Sen, 1970); however, they stay independent. The convergence in probability was studied in (Mizera and Wellner, 1998), where necessary and sufficient conditions for the convergence of the sample median were discovered. As it is clear from (2.36) and (2.37) the assumption of independence is not fulfilled in our case.

Example 2.2.11. *Let us consider the time series $y_t = a_{0,t} + a_{1,t}t + \varepsilon_t$ for $t = 101, \dots, n$ where $a_{0,t} = 0$, $a_{1,t} = 1$, ε_t has a normal distribution* $\mathrm{N}(0,1)$ *for all $t = 101, \dots, n$. We further suppose that our initial estimates $\tilde{\boldsymbol{a}}_{101}$ are misleading namely $\tilde{a}_{0,101} = 100$ and $\tilde{a}_{1,101} = -100$. We do not check the statistics in Step 14 of Algorithm 2.2.2 i.e., we proceed the loop until $t = n$, put $\hat{y}_t = \hat{a}_{0,t} + \hat{a}_{1,t}t$ for $t = 101, \dots, n$ and $W = 3$. Under these conditions we get the following results for a simulated series.*

Table 2.2. Results of Algorithm 2.2.2 without evaluation of A^t_{101+k} for $k = t - 101, \dots, 1$ applied on a time series where $a_{0,t} = 0$, $a_{1,t} = 1$, $\tilde{a}_{0,101} = 100$, $\tilde{a}_{1,101} = -100$ and ε_t has distribution $\mathrm{N}(0,1)$.

t	$\bar{a}_{0,t}$	$\bar{a}_{1,t}$	$\hat{a}_{0,t}$	$\hat{a}_{1,t}$	$a_{0,t} + a_{1,t}t - \hat{y}_t$
$t = 104$	100	-98	10251	0.01	-10148
$t = 110$	5297	-46	5225	-46	47
$t = 200$	7680	-31	6537	-37	1126
$t = 500$	12833	-18	9768	-24	3053
$t = 1000$	18504	-12	13510	-17	4984

We present also a table, where $\tilde{\boldsymbol{a}}_{101}$ is reasonable.

Table 2.3. Results of Algorithm 2.2.2 without evaluation of A^t_{101+k} for $k = t - 101, \dots, 1$ applied on a time series where $a_{0,t} = 0$, $a_{1,t} = 1$, $\tilde{a}_{0,101} = 1.1$ and $\tilde{a}_{1,101} = 0.2$ and ε_t has distribution $\mathrm{N}(0, 1)$.

t	$\bar{a}_{0,t}$	$\bar{a}_{1,t}$	$\hat{a}_{0,t}$	$\hat{a}_{1,t}$	$a_{0,t} + a_{1,t}t - \hat{y}_t$
$t = 104$	1,711	1,11	-10,53	0,99	11,07
$t = 110$	-5,44	1,04	-5,34	1,04	0,62
$t = 200$	-8,37	1,03	-6,19	1,04	-1,16
$t = 500$	-9,89	1,02	-9,19	1,02	-3,01
$t = 1000$	-17,09	1,01	-12,92	1,02	-4,25

We can make a conjecture with respect to these results that it is sufficient to compute $\tilde{\boldsymbol{a}}_{t_j}$ and not to perform Steps 11 and 12 of Algorithm 2.2.2. We can still see the advantage of our method, because the estimate of a_1 is improving with a number of observations.

△

It is visible from Example 2.2.11 that Algorithm 2.2.2, under the conditions of the example, generally does not converge and that it is highly dependent on the initial estimate $\tilde{\boldsymbol{a}}_{t_j}$. On the other hand if $\tilde{\boldsymbol{a}}_{t_j}$ is misspecified then $|\hat{a}_{0,t} + \hat{a}_{1,t}t - y_t|$ starts to grow and the difference $\hat{a}_{0,t} + \hat{a}_{1,t}t - y_t$ does not switch the sign for t high enough. This is visible from (2.36) and (2.37), because $\delta_{0,t}$ and $\delta_{1,t}$ have the same sign if $|\delta_{i,t}| \gg |\varepsilon_t|$. The growth of $|\hat{a}_{0,t} + \hat{a}_{1,t}t - y_t|$ leads to fulfillment of the condition in Step 14 of Algorithm 2.2.2. I.e., if the initial estimate is misspecified then we find very early a change point and if the next initial estimate $\tilde{\boldsymbol{a}}_{t_{j+1}}$ is already correctly specified then not too many observations are wrongly estimated.

Lemma 2.2.12. *Let us have observations $(y_1, \dots, y_n)^\top$, $W \in \mathbb{N}$, $b > 0$, $\tilde{a}_{0,t}$ is shift and scale equivariant, $\tilde{a}_{1,t}$ is shift invariant and scale equivariant for $t = 1, \dots, n$. We define $\hat{y}_t$ according to Algorithm 2.2.2, then $\hat{y}_t$ for $t = 1, \dots, n$ is shift and scale equivariant.*

Proof. We proceed by the mathematical induction. Let us have a change point in time $t_j \in \{1, \dots, n\}$. We know from assumptions that $\tilde{a}_{0,t}$ is shift and scale equivariant and that $\tilde{a}_{1,t}$ is shift invariant and scale equivariant. It yields that $\hat{a}_{0,t_j+W}$ is also shift and scale equivariant and $\hat{a}_{1,t_j+W}$ is shift invariant and scale equivariant. We get from equations 2.33 and 2.34 for $\tau = t_j + 1, \dots, t_j + W$ that $\bar{a}_{0,\tau}(\hat{\boldsymbol{a}}_{\tau-1})$ is shift and scale equivariant and $\bar{a}_{1,\tau}(\hat{\boldsymbol{a}}_{\tau-1})$ is shift invariant and scale equivariant. This is also valid for $\bar{a}_{0,t_j}(\hat{\boldsymbol{a}}_{t_j})$ and $\bar{a}_{1,t_j}(\hat{\boldsymbol{a}}_{t_j})$.

We suppose for $n \geq t > t_j + W$ that no change point is found between t_j and t and that $\hat{a}_{0,t-1}$ is shift and scale equivariant and $\hat{a}_{1,t-1}$ is shift invariant and scale equivariant. We further suppose that $\bar{a}_{0,t-1}(\hat{\boldsymbol{a}}_{t-2})$ is shift and scale equivariant and $\bar{a}_{1,t-1}(\hat{\boldsymbol{a}}_{t-2})$ is shift invariant and scale equivariant. We want to show that $\hat{a}_{0,t}$ is shift and scale equivariant and $\hat{a}_{1,t}$ is shift invariant and scale equivariant. We employ once more equations 2.33 and 2.34 to get that $\bar{a}_{0,t}(\hat{\boldsymbol{a}}_{t-1})$ is shift and scale equivariant and $\bar{a}_{1,t}(\hat{\boldsymbol{a}}_{t-1})$ is shift invariant and scale equivariant. For median is shift and scale equivariant and $\hat{\boldsymbol{a}}_t$ is computed as a median from $\bar{\boldsymbol{a}}_{t_j}(\hat{\boldsymbol{a}}_{t_j}), \bar{\boldsymbol{a}}_{t_j+1}(\hat{\boldsymbol{a}}_{t_j}), \ldots, \bar{\boldsymbol{a}}_{t-1}(\hat{\boldsymbol{a}}_{t-2}), \bar{\boldsymbol{a}}_t(\hat{\boldsymbol{a}}_{t-1})$ we get that $\hat{a}_{0,t}$ is shift and scale equivariant and $\hat{a}_{1,t}$ is shift invariant and scale equivariant.

Let $\bar{y}_t = cy_t + s$ for $t = 1, \ldots, n$ and $c, s \in \mathbb{R}$. We get

$$\begin{aligned}\hat{\bar{y}}_t &= \hat{\boldsymbol{a}}_{0,t}((\bar{y}_1, \ldots, \bar{y}_n), b, W) + \hat{\boldsymbol{a}}_{1,t}((\bar{y}_1, \ldots, \bar{y}_n), b, W)t \\ &= c\hat{\boldsymbol{a}}_{0,t}((y_1, \ldots, y_n), b, W) + s + c\hat{\boldsymbol{a}}_{1,t}((y_1, \ldots, y_n), b, W)t \\ &= c\hat{y}_t + s.\end{aligned}$$

We can follow the proof of Lemma 2.2.3 for the rest of the proof. □

We leave the problematic of breakdown point to Section 2.2.4, because the initial algorithm of this section based on absolute values is not robust in the sense of breakdown point. If we utilize another algorithm such as the least weighted squares then we would get similar results as in the following section.

We want to discuss now a time complexity see Definition 1.1.13 of Algorithm 2.2.2. We consider the worst case as in the case of constant trend, meaning that no level shifts are found. I.e., we have to compute for each t the median $\boldsymbol{M}_t\left(1, (\bar{\boldsymbol{a}}_1(\hat{\boldsymbol{a}}_1), \bar{\boldsymbol{a}}_2(\hat{\boldsymbol{a}}_1), \ldots, \bar{\boldsymbol{a}}_{t-1}(\hat{\boldsymbol{a}}_{t-2}), \bar{\boldsymbol{a}}_t(\hat{\boldsymbol{a}}_{t-1}))\right)$. We can once more exploit the information about ordering from the previous step and therefore we get in each step the time complexity $\mathcal{O}(2t)$. We multiply by 2 due to computing two medians. It holds in the last step $n = t$, therefore we replace t by n and get the time complexity of the algorithm $\mathcal{O}(2n^2)$. The constant can be neglected so the time complexity is $\mathcal{O}(n^2)$.

We should note that in the general case of higher dimension d it would be necessary to compute more medians for each t, and therefore the complexity would grow to $\mathcal{O}(dn^2)$, which is still only $\mathcal{O}(n^2)$.

The time complexity depends on b in reality. If b is small then we have to order only a few observations and so the complexity is close to $\mathcal{O}(n)$.

2.2.4 Sign test: Linear trend - modification

We deal with a modification of Algorithm 2.2.2 in this part such that there is higher potential of better properties in the sense of convergence than Algorithm 2.2.2. For the discussion about convergence of Algorithm 2.2.2 see Example 2.2.11. I.e., we try to propose an algorithm for which the addition of a new observation can improve the final estimate and for which the final estimate is not so dependent on the initial estimates. The modification is based on the idea described in (Siegel, 1982) known as repeated median algorithm.

The notation in this part is the same as in 2.2.3. Some differences are described below. We define for some $t, \tau \in \mathbb{N}$ and $\tau < t \leq n$.

$$\breve{a}_{1,t}(\tau) = \frac{y_t - y_\tau}{t - \tau}. \tag{2.38}$$

We denote a median from observations $c_{t_j}, \ldots, c_t$ as $M(c_{t_j}, \ldots, c_t)$.

We describe in the following algorithm how to gain an initial estimate $\tilde{\boldsymbol{a}}_t$. We employ a similar procedure as in the case of the final estimate $\hat{\boldsymbol{a}}_t$ (compare 2.2.5) with the difference that we do not compute the statistics $A^t_{t_j+k}(\hat{a}_{0,t} + \hat{a}_{1,t}t)$ for $k = t - t_j, \ldots, 1$. We also gain $\bar{\boldsymbol{a}}_\tau$ for $\tau = t+1, \ldots, t+W$ during the computation of $\tilde{\boldsymbol{a}}_t$ which are later utilized in Algorithm 2.2.5.

Algorithm 2.2.4 Computation of $\tilde{\boldsymbol{a}}_t$

```
**Input:** time $t$, $W \in \mathbb{N}$, observations $(y_t, \dots, y_{t+W})^\top$
$\bar{a}_{0,t} \leftarrow y_t$
**for** $\tau = t+1, \dots, t+W$ **do**
    compute $\check{a}_{1,\tau}(i)$ for $i = t, \dots, \tau - 1$
    $\bar{a}_{1,\tau} \leftarrow M(\check{a}_{1,\tau}(t), \dots, \check{a}_{1,\tau}(\tau - 1))$
    $\hat{a}_{1,\tau} \leftarrow M(\bar{a}_{1,t+1}, \dots, \bar{a}_{1,\tau})$
    $\bar{a}_{0,\tau} \leftarrow y_\tau - \tilde{a}_{1,\tau}(\tau - t)$
    $\hat{a}_{0,\tau} \leftarrow M(\bar{a}_{0,t+1}, \dots, \bar{a}_{0,\tau})$
**end for**
$\tilde{a}_{0,t} \leftarrow M(\bar{a}_{0,t}, \dots, \bar{a}_{0,t+W})$
$\tilde{a}_{1,t} \leftarrow \hat{a}_{1,t+W}$
```

We present the main part of the algorithm based on repeated median estimator.

Algorithm 2.2.5 Algorithm for a linear trend based on the sign test and repeated median estimator

```
**Input:** observations $(y_1, \dots, y_n)^\top$, $W \in \mathbb{N}$, $b > 0$
put $t \leftarrow 1$, $j \leftarrow 1$, $t_1 \leftarrow 1$
**while** $t \leq n$ **do**        ▷ this has to be checked throughout the While loop
    **if** $W \geq 2$ **then**
        compute $\tilde{\boldsymbol{a}}_{t_j}$ according to 2.2.4 with $W = W - 1$
    **end if**
    $t \leftarrow t_j + W - 1$
    **repeat**
        $t \leftarrow t + 1$
        compute $\check{a}_{1,t}(\tau)$ for $\tau = t_j, \dots, t - 1$
        $\bar{a}_{1,t} \leftarrow M(\check{a}_{1,t}(t_j), \dots, \check{a}_{1,t}(t - 1))$
        $\hat{a}_{1,t} \leftarrow M(\bar{a}_{1,t_j+1}, \dots, \bar{a}_{1,t})$
        $\bar{a}_{0,t} \leftarrow y_t - \hat{a}_{1,t}(t - t_j)$
        $\hat{a}_{0,t} \leftarrow M(\bar{a}_{0,t_j}, \dots, \bar{a}_{0,t})$
        compute $A^t_{t_j+k}(\hat{a}_{0,t} + \hat{a}_{1,t}t)$ for $k = t - t_j, \dots, 1$
    **until** $A^t_{t_j+k}(\hat{a}_{0,t} + \hat{a}_{1,t}\tau) \in (-b, b)$ for $k = t - t_j, \dots, 1$
```

```
16:     $j \leftarrow j+1$
17:     $t_j \leftarrow t_{j-1} + \min\left\{k \middle| \left|A^t_{t_j+k}(\hat{a}_{0,t} + \hat{a}_{1,t}\tau)\right| > b; k = t_j - t, \dots, 1\right\}$
18:     compute $\tilde{\boldsymbol{a}}_{t_j}$ according to 2.2.4
19:     $m \leftarrow -1$
20:     while $|\tilde{a}_{0,t_j} + \tilde{a}_{1,t_j}m - y_{t_j-1}| < |\hat{a}_{0,t_j} + \hat{a}_{1,t_j}(t_j - 1 - t_{j-1}) - y_{t_j-1}|$ and $t_j > t_{j-1}$ do
21:         $t_j \leftarrow t_j - 1$
22:         $m \leftarrow m - 1$
23:     end while
24:     while $|\tilde{a}_{0,t_j} + \tilde{a}_{1,t_j} - y_{t_j+1}| > |\hat{a}_{0,t_j} + \hat{a}_{1,t_j}(t_j + 1 - t_{j-1}) - y_{t_j+1}|$ and $t_j < t$ do
25:         $t_j \leftarrow t_j + 1$
26:         compute $\tilde{\boldsymbol{a}}_{t_j}$ according to 2.2.4
27:     end while
28:     if $t_j - t_{j-1} < W$ then
29:         $\hat{y}_{t_{j-1}}, \dots, \hat{y}_{t_j-1} \leftarrow \tilde{a}_{0,t_{j-1}}(y_{t_{j-1}}, \dots, y_{t_j-1}) + \tilde{a}_{1,t_{j-1}}(y_{t_{j-1}}, \dots, y_{t_j-1})t$
30:     else
31:         $\hat{y}_{t_{j-1}}, \dots, \hat{y}_{t_j-1} \leftarrow \hat{a}_{0,t_j} + \hat{a}_{1,t_j}t$
32:     end if
33: end while
```

We put $\bar{a}_{0,t_j} = y_{t_j}$ in Algorithm 2.2.5 Step 12 and Procedure 2.2.4 Step 6. I.e., we consider each change point as a new beginning of time. It has the advantage that we can employ one more observation in the computation of $\hat{a}_{0,t}$ which is independent from $\hat{a}_{1,t}$. This is not necessary in the case of Algorithm 2.2.2, because we use a different initial estimate.

If we want to utilize more information from our data, we can apply the whole procedure of the repeated median algorithm in each step. It means we have to adjust Steps 9 and 10 of Algorithm 2.2.5 in the following way.

Algorithm 2.2.6 Possible replacement of Steps 9 and 10 of Algorithm 2.2.5

```
 7: compute $\check{a}_{1,\tau}(t)$ for $\tau = t_j, \dots, t-1$
 8: compute $\check{a}_{1,t}(\tau)$ for $\tau = t_j, \dots, t-1$
 9: for $\tau = t_j, \dots, t$ do
10:     $\bar{a}_{1,\tau} \leftarrow M(\check{a}_{1,\tau}(t_j), \dots, \check{a}_{1,\tau}(\tau-1), \check{a}_{1,\tau}(\tau+1), \dots, \check{a}_{1,\tau}(t))$
11: end for
```

It is also necessary to adjust initial Procedure 2.2.4. This modification demands more computational effort.

Lemma 2.2.13. *Let us have observations $(y_1, \dots, y_n)^\top$, $W \in \mathbb{N}$ and $b > 0$. We define $\hat{y}_t$ according to Algorithm 2.2.5, then $\hat{y}_t$ for $t = 1, \dots, n$ is shift and scale equivariant.*

Proof. We can follow the proof of Lemma 2.2.12. It suffices to show, in the case of this lemma, shift invariance and scale equivariance of $\check{a}_{1,t}(\tau)$, shift invariance and scale equivariance of $\bar{a}_{1,t}$ and shift and scale equivariance of $\bar{a}_{0,t}$ for $t, \tau =$

$1, \dots, n$ and $\tau < t$. It follows from (2.38) that $\breve{a}_{1,t}(\tau)$ is shift invariant and scale equivariant. The shift invariance and scale equivariance of $\bar{a}_{1,t}$ follows from the fact that it is a median from $\breve{a}_{1,t}(\tau)$ for appropriate values of t and value τ. It follows already from that the shift invariance and scale equivariance of $\hat{a}_{1,t}$. Since $\bar{a}_{0,t} = y_t - \hat{a}_{1,t}(t-\tau)$, we get that $\bar{a}_{0,t}$ is shift and scale equivariant.

□

We should note that the choice of the repeated median algorithm was not too fortunate, because its asymptotic properties do not represent a big improvement with respect to the previous section. According to (Siegel, 1982), it is Fisher consistent 1.1.5 and unbiased. Its convergence was also studied in (Hossjer et al., 1994). Nevertheless, its asymptotic properties are not convincing. Therefore, we should not await great improvement. On the other hand this algorithm can serve as a good starting point for further modifications. One of the possible approaches could be to employ a robust version of Kalman filter see e.g., (Hanzák and Cipra, 2011).

We can prove a similar assertion to 2.2.4. I.e., we have to consider once more the advantages and disadvantages of a size of b. However, we cannot say that for high values of b we get converging estimators and therefore closer to the real value.

If we want to study breakdown point we have to employ Modification 2.2.6 to gain exactly the same algorithm which was described in (Siegel, 1982). Let T_n denote the estimator of the time series given by Algorithm 2.2.5 with modification 2.2.6. We employ the notation from Definition 1.1.7. From (Siegel, 1982) we know that $\varepsilon^*_{m(t,\boldsymbol{y})}(T_{n,t}, \boldsymbol{z}_t) = \frac{1}{m(t,\boldsymbol{y})} \frac{m(t,\boldsymbol{y})-2}{2}$.

Example 2.2.14. *We show in this example, why the Algorithm 2.2.5 without modification 2.2.6 is not as robust as with the modification. Consider observations* $y_1 = 1, y_2 = 2, y_3 = 0, y_4 = 0, y_5 = 0, y_6 = 0$ *then* $\bar{a}_{1,2} = 1, \bar{a}_{1,3} = -1.25, \bar{a}_{1,4} = -\frac{1}{3}, \bar{a}_{1,5} = -\frac{1}{8}, \bar{a}_{1,6} = 0$. *Therefore, we spoiled four* $\bar{a}_{1,t}$ *for* $t = 2, \dots, 6$ *by contaminating the first two observations. It is easy to see that this holds generally. If we contaminate the observations in the beginning we get twice as many contaminated observations* $\bar{a}_{1,t}$. *We need to have more observations* $\bar{a}_{1,t}$ *uncontaminated to get an unviolated final estimate. It yields approximately the breakdown point of Algorithm 2.2.5 to be one fourth.* △

Remark 2.2.15. *We have to take into account Steps 20-23 of Algorithm 2.2.5 to study similar number as* $k_{b,t}$ *from* (2.22). *These steps can shrink possible minimal number of observations between two change points. Therefore, there is no need to proof similar proposition as 2.2.7. As the second counterargument serves the fact that we cannot exploit the assumption that half of observations lie below the estimated values and half of observations lie above.*

We get according to Definition 1.1.10 with respect to $\varepsilon^*_{m(t,\boldsymbol{y})}(T_{n,t}, \boldsymbol{z}_t) = \frac{1}{m(t,\boldsymbol{y})} \frac{m(t,\boldsymbol{y})-2}{2}$ that

$$\pi_t(T_n(\bar{\boldsymbol{y}})) = \sum_{i=0}^{\left\lfloor \frac{m(t,\boldsymbol{y})-2}{2} \right\rfloor} p^i (1-p)^{m(t,\bar{\boldsymbol{y}})-i}.$$

We will discuss the time complexity of Algorithm 2.2.5. We have to compute median $M(\check{a}_{1,t}(t_j),\ldots,\check{a}_{1,t}(t-1))$ for each t in comparison to Algorithm 2.2.2 and cannot employ the ordering from the previous step. Therefore, we get the complexity $\mathcal{O}(n^2 \log n)$ based on the fact that the complexity of the ordering is at least $\mathcal{O}(n \log n)$ see (Cormen, 2009)[151].

We can replace Steps 20 23 of Algorithm 2.2.5 in a similar way as in the case of modification 2.2.3 of Algorithm 2.2.2. However, the modification does not lead to an improvement in this case and it slows down the algorithm. Therefore, we does not employ it in the simulation study.

2.3 Simulation study

In this Section, simulations for models with constant and linear trends are presented with the aim to find the optimal arrangement of the corresponding procedures. We also compare different methods.

2.3.1 Simulation study: Constant trend

We have generated, in Matlab and C, time series of length $n = 100$ with a constant trend $y_t = a + \varepsilon_t$. The errors are i.i.d. $\mathrm{N}(0,1)$ but they are, with probability p, contaminated by other distributions specified in Tables 2.4, 2.5 and 2.6 (e.g., $\mathrm{N}(0,100)$ with probabilities $p = 5\%$ or $p = 10\%$). Compare with Definition 1.1.2. We have always generated $N = 1000$ series of the same type for particular situations.

Let a_t denote the actual value of a at time t and let $\widehat{y}_t$ be the estimate of a_t (i.e., the smoothed value) based on one of the compared algorithms. If we also want to stress the series number i, we add the index i to the relevant symbol, e.g., $y_{t,i}$.

Moreover, the level shift occurs at time $t = 50$. In particular, for each time series the values a_t for $t = 1,\ldots,49$ are constant, generated by the uniform distribution on the interval (-10,10) for each trajectory and the same rule holds for $t = 50,\ldots,100$ (all samples are independent).

There is a criterion MAE (Mean Absolute Error) to be minimized with respect to the technical coefficients b and W from Algorithm 2.2.1

$$\mathrm{MAE} = \frac{1}{Nn}\sum_{i=1}^{N}\sum_{t=1}^{n}|a_{t,i} - \widehat{y}_{t,i}|.$$

We also try to employ the technical coefficients in order to minimize MAE for other algorithms in this simulation. The MAE criterion's value differ for the case of forecasting (see later).

We look for b and W such that they minimize the criterion MAE in Table 2.4. If the criterion or a coefficient is indexed by min, then its value has been obtained within minimization of MAE.

We compare the values of $\mathrm{MAE}_{\mathrm{min}}$ with the values obtained under the condition that the technical coefficients b and W are fixed (then the objective function is not indexed and the values of coefficients are given in the legend of the corresponding table).

With respect to the first column and other computations it turns out that the results do not depend too much on the technical coefficient W. More precisely, for a wide range of values of W, the results are almost the same. Therefore, for the sign test algorithm with the constant trend, one can recommend the choice of this value between 35 and 50. But we should also take into account the approximate length between two consecutive change points. I.e., the optimal W can depend on the distance between change points. The boundary b can be recommended around 3 (on the other hand, the results are not too dependent on this value and we can choose any value between 2.5 and 3.5). The values for b are appropriate, especially for smoothing. We should recommend lower values (between 1.5 and 2) for forecasting, because we could forecast incorrectly after a level shift. A recommendation for routine application of the sign test algorithm with constant trend follows: b equal to 3 and W higher than 20 or equal to the average conjectural length between two change points.

Table 2.4. (constant trend by the sign test algorithm) There are values of MAE for fixed $b = 2$ and $W = 50$ in the first column. There are the minimal values of MAE in the second column. There are also values of technical coefficients $b_{\min}$ and $W_{\min}$ minimizing MAE for the specified series in the third and fourth columns.

Distribution	MAE	MAE$_{\min}$	$b_{\min}$	$W_{\min}$
$p = 0\%$	0.164	0.157	2.760	49
$p = 5\%$				
N$(0, 100)$	0.198	0.168	3.159	50
Cauchy	0.173	0.155	2.904	50
U$(-10, 10)$	0.190	0.173	2.866	50
$p = 10\%$				
N$(0, 100)$	0.217	0.191	2.935	50
Cauchy	0.181	0.164	3.007	49
U$(-10, 10)$	0.215	0.187	3.135	45
$p = 40\%$				
U$(-20, 20)$	2.779	0.380	0.453	55

Let us compare now the sign test algorithm with other algorithms. It is also interesting to see the average size of residuals, because it can help us to decide whether the estimate given by a specific algorithm really smoothes the values of the series y_t in the direction of a_t. In other words, whether the estimate is closer to a_t than the original series y_t. For this purpose we employ the average of absolute errors

$$\frac{1}{Nn}\sum_{i=1}^{N}\sum_{j=1}^{n}|a_{t,i} - y_{t,i}|.$$

We employ the following indices:

(i) average of absolute errors (Err);

(ii) simple exponential smoothing (index Exp);

(iii) C-algorithm (index C);

(iv) sign test algorithm (index S);

(v) M-estimation of simple exponential smoothing (index M).

E.g., MAE_{Exp} means that we substitute the estimates from the exponential smoothing algorithm to the MAE criterion. The same technical coefficients $(\beta_{Exp}, \beta_C, b, W)$ are applied to each kind of outliers so their choice is rough.

The M-estimation method is adopted from (Hanzák and Cipra, 2011). Under the assumption, that the level shifts appear quite rarely, the method (Hanzák and Cipra, 2011) gives moderate results. On the other hand, when we generate the data according to the article, then their robust method gives better results than the method described here. Obviously, it depends on the nature of the data:

1. If we suppose a small change in each step then M-estimation should be preferred.

2. In the case of significant rare jumps one should employ the method from our work.

Table 2.5 shows that according to the MAE criterion, the sign test algorithm works best by a significant margin.

Table 2.5. (constant trend smoothing: comparison of algorithms) (i) the average of absolute errors of simulated time series; the values of MAE for (ii) the simple exponential smoothing with $\beta_{Exp} = 0.6$; (iii) the C-algorithm with $\beta_C = 0.6$; (iv) the sign test with $b = 2$ and $W = 50$; (v) M-estimation of simple exponential smoothing with $\alpha_M = 0.8$ and $\nu_M = 0.5$.

Distribution	Err	MAE_{Exp}	MAE_C	MAE_S	MAE_M
$p = 0\%$	0.797	0.487	0.603	0.164	0.583
$p = 5\%$					
$\text{N}(0, 100)$	1.144	0.752	0.654	0.198	0.600
Cauchy	1.020	0.685	0.613	0.173	0.579
$\text{U}(-10, 10)$	1.007	0.635	0.648	0.190	0.609
$p = 10\%$					
$\text{N}(0, 100)$	1.513	1.017	0.730	0.217	0.637
Cauchy	1.468	1.108	0.632	0.181	0.586
$\text{U}(-10, 10)$	1.217	0.771	0.708	0.215	0.639
$p = 40\%$					
$\text{U}(-20, 20)$	4.434	2.790	2.112	0.390	1.462

The sign test method smoothes a value of the series based on its current, past and future values. However, the exponential smoothing based methods, employed for comparison in our simulation study, use just current and past values of the series. Thus, in the case of smoothing, the comparison is not "fair".

Let us turn our attention to forecasting. If we want to forecast, then the crucial thing is to detect the level shift faster than in the case of smoothing, because after the level shift it may happen that we predict too many observations incorrectly (according to old observations). We employed the parameter b equal to 2 and not higher also in the case of smoothing, to keep comparable conditions for all algorithms.

In the case of forecasting, we employ the following function

$$\mathrm{MAE}_f = \frac{1}{N(n-10)} \sum_{i=1}^{N} \sum_{t=10}^{n-1} |a_{t+1,i} - \widehat{y}_{t+1,i|t}|,$$

where $\widehat{y}_{t+1,i|t}$ stands for an estimate of $a_{t+1,i}$, if we know the observations $y_1, \dots, y_t$.

We see from Table 2.6 that our algorithm is still the best, but it is much closer to the others. The values of MAE_f are generally worse. This is given by the fact that for prediction it is harder to switch to another level after a level shift, and the prediction is always delayed.

Table 2.6. (constant trend forecasting: comparison of algorithms) (i) the average of absolute errors of simulated time series; the values of MAE_f for (ii) the simple exponential smoothing with $\beta_{Exp} = 0.6$; (iii) the C-algorithm with $\beta_C = 0.6$; (iv) the sign test with $b = 2$ and $W = 50$; (v) M-estimation of simple exponential smoothing with $\alpha_M = 0.8$ and $\nu_M = 0.5$.

Distribution	Err	$\mathrm{MAE}_{f,Exp}$	$\mathrm{MAE}_{f,C}$	$\mathrm{MAE}_{f,S}$	$\mathrm{MAE}_{f,M}$
$p = 0\%$	0.797	0.564	0.675	0.560	0.672
$p = 5\%$					
N(0, 100)	1.150	0.835	0.724	0.575	0.699
Cauchy	1.136	0.894	0.681	0.552	0.673
U(−10, 10)	1.010	0.710	0.716	0.575	0.701
$p = 10\%$					
N(0, 100)	1.514	1.094	0.794	0.596	0.738
Cauchy	1.268	0.958	0.699	0.562	0.686
U(−10, 10)	1.226	0.848	0.771	0.581	0.723
$p = 40\%$					
U(−20, 20)	4.434	2.840	2.121	0.924	1.542

We present a time series with large residuals and also a level shift in Figure 2.2. We can visually compare in this picture how the algorithms are able to deal with these obstacles. For instance, the M-algorithm deals quite well with high residuals but is unable to react fast enough to a quite large level shift. The C-algorithm and the exponential smoothing algorithm behave in a similar way; nevertheless the exponential smoothing is more sensitive to violations in data.

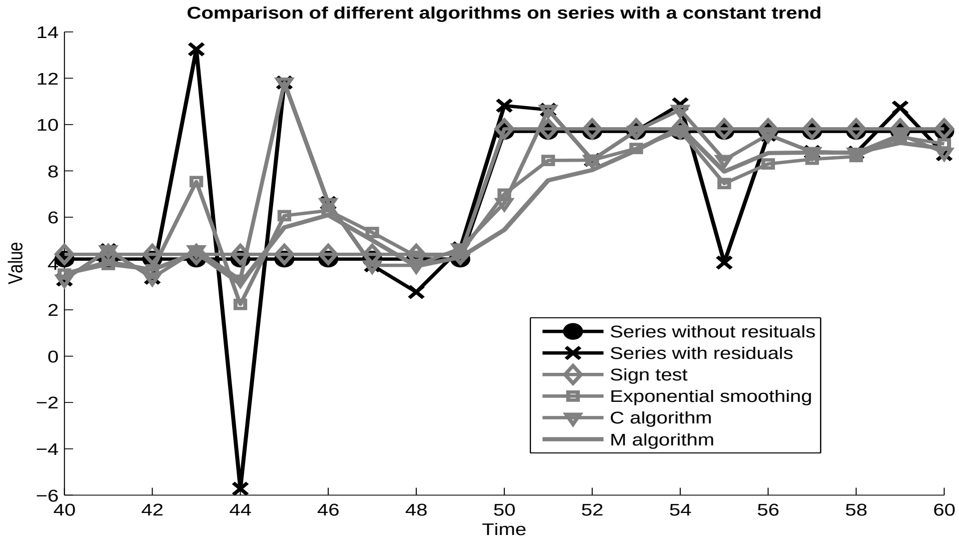


Figure 2.2. (constant trend): Comparison of different estimates of time series ($\alpha = 0.5, \beta_C = 0.6, \beta_{Exp} = 0.6, W = 50, b = 2, \alpha_M = 0.8$ and $\nu_M = 0.5$).

2.3.2 Simulation study: Linear trend

The situation differs from the case of the constant trend in the linear case, since we also have to deal with a slope. It is possible to generate the series by many methods. We choose only one here. The notation is the same as above (e.g., MAE $= \frac{1}{Nn}\sum\limits_{i=1}^{N}\sum_{t=1}^{n}|\beta_{0,i} + \beta_{1,i}t - \widehat{y}_{t,i}|$).

Similar to the previous case, we generate the time series of length $n = 100$, and $N = 100$ series for each distribution. The number N differs from the case of the constant trend, because the computation in the linear case can already be quite time-consuming if $N = 1000$, especially when we want to minimize and look for the most suitable technical coefficients.

The constant term is initialized by the uniform distribution on the interval $\langle -10; 10\rangle$ and the linear term similarly on the interval $\langle -5; 5\rangle$. The constant and slope coefficients are changed at time $t = 50$ for each time series. The new linear term also follows the uniform distribution on the interval $\langle -5; 5\rangle$. To avoid too large a gap between observations 49 and 50, we put the constant term β_0 for the new line such that $a_{50} = a_{49} + u$, where u has the uniform distribution on the interval $\langle -10; 10\rangle$. The residuals are constructed in the same way as for a constant trend. We have to also deal with the technical coefficients W and b. We employ the same methods as in the case of a constant trend.

We employ the version of the sign test algorithm following option b (see paragraph 2.2.3), because the algorithm following Remark (a) does not work properly in some cases.

Let us look now at Table 2.7, where we display the optimal technical coefficients. One can recommend a choice of the window W in a length equal to the probable length between two neighboring level shifts for the linear trend by the sign test algorithm. The coefficient b should be chosen between 2.2 and 2.5. Once more, the choice of technical coefficients is not too important, because we get very similar results for quite a wide range of values (see Table 2.7).

Table 2.7. (linear trend by the sign test algorithm) This Table employs Algorithm 2.2.2 including the Steps 17-24. There are values of MAE for fixed $b = 2$ and $W = 50$ in the first column. There are the minimal values of MAE in the second column. There are also the values of technical coefficients $b_{\min}$ and $W_{\min}$ minimizing MAE in the third and fourth columns.

Distribution	MAE	MAE$_{\min}$	$b_{\min}$	$W_{\min}$
$p = 0\%$	0.231	0.219	2.293	53
$p = 5\%$				
$\mathrm{N}(0, 100)$	0.289	0.250	2.520	52
Cauchy	0.271	0.221	2.286	53
$\mathrm{U}(-10, 10)$	0.244	0.226	2.222	52
$p = 10\%$				
$\mathrm{N}(0, 100)$	0.299	0.291	2.218	50
Cauchy	0.343	0.212	2.437	51
$\mathrm{U}(-10, 10)$	0.265	0.241	2.220	53
$p = 40\%$				
$\mathrm{U}(-20, 20)$	0.640	0.594	1.894	50

We employ the following indices from now:

(i) average of absolute errors (Err);

(ii) double exponential smoothing (index Exp);

(iii) linear sign test algorithms without looking for change points in Steps 17 - 24 of Algorithm 2.2.2 (index SL_b);

(iv) linear sign test algorithm (index SLT_b);

(v) modified linear sign test Algorithm 2.2.5 (index STS);

(vi) double exponential smoothing employing M-estimation (index M).

We can compare the sign test algorithm with the double exponential smoothing. The coefficient was chosen to be optimal according to MAE and with respect to (Brown, 1962), where its author recommends an interval for values of the coefficient for the double exponential smoothing method. Our algorithms give significantly better results in the contaminated cases.

We also employ M-estimation of double exponential smoothing and compare it with other algorithms in Table 2.8. The detailed description of this method can be found in (Hanzák and Cipra, 2011). The results for M-estimation are better for a higher contamination in (relative) comparison to other methods (similarly as for the constant trend): this approach should be used in the case of small jumps occurring quite often since the M-estimation algorithm has been originally suggested for such a type of data. Otherwise, we should employ our algorithm.

We can also see from the table that modified Algorithm 2.2.5 does not lead to the improvement. It is generally worse than original sign test Algorithm 2.2.2. This could be caused by the slow convergence of the repeated median algorithm, because for an initiation of the original sign test algorithm we utilize the absolute regression, which is faster. This is more obvious in our situation when $W = 50$. However, as was already mentioned further research in this way is necessary.

Table 2.8. (linear trend smoothing: comparison of algorithms) (i) the errors i.e., $Err = \frac{1}{Nn}\sum_{i=1}^{N}\sum_{j=1}^{n}|a_{t,i} - y_{t,i}|$; (ii) the double exponential smoothing with $\beta = 0.84$ (Exp); (iii) the linear sign test algorithms without the improvement, connected with looking for change points from Steps 17-24 of Algorithm 2.2.2 ($W = 10$, $b = 2$); (iv) the linear sign test algorithm ($W = 50$, $b = 2$), considering now the improvement; (v) the modified linear sign test Algorithm 2.2.5 ($W = 50$, $b = 3$); (vi) the double exponential smoothing employing M-estimation with parameters $\alpha_M = 0.8$ and $\nu_M = 0.7$.

Distribution	Err	MAE$_{Exp}$	MAE$_{SL_b}$	MAE$_{SLT_b}$	MAE$_{STS}$	MAE$_M$
$p = 0\%$	0.797	0.722	0.513	0.224	0.337	2.808
$p = 5\%$						
N$(0, 100)$	1.168	1.076	0.546	0.271	0.361	2.761
Cauchy	0.975	0.906	0.516	0.239	0.334	2.743
U$(-10, 10)$	1.011	0.917	0.549	0.248	0.354	2.455
$p = 10\%$						
N$(0, 100)$	1.512	1.397	0.597	0.323	0.401	2.770
Cauchy	1.299	1.251	0.526	0.260	0.348	3.193
U$(-10, 10)$	1.216	1.097	0.578	0.263	0.376	2.867
$p = 40\%$						
U$(-20, 20)$	4.469	3.969	1.325	0.640	0.753	4.734

We study the ability of algorithms to forecast similar to the case of a constant trend. We see from Table 2.9 that our algorithm is the best in almost all cases except for the non-contaminated case, but we have to employ the improved algorithm. It is more suitable to use lower b around 1.5 in the case of forecasting. We get even better results then. We omitted the modified sign test algorithm for forecasting.

Table 2.9. (linear trend forecasting: (i) the errors i.e., $Err = \frac{1}{Nn}\sum_{i=1}^{N}\sum_{j=1}^{n}|a_{t,i} - y_{t,i}|$; (ii) the double exponential smoothing with $\beta = 0.84$ (Exp); (iii) the linear sign test algorithms without the improvement, connected with looking for change points from Steps 17-24 of Algorithm 2.2.2 ($W = 10$, $b = 2$); (iv) the linear sign test algorithm ($W = 50$, $b = 2$), considering now the improvement; (vi) the double exponential smoothing employing M-estimation with parameters $\alpha_M = 0.8$ and $\nu_M = 0.7$.

Distribution	Err	MAE$_{f,Exp}$	MAE$_{f,SL_b}$	MAE$_{f,SLT_b}$	MAE$_{f,M}$
$p = 0\%$	0.796	0.979	1.589	1.036	3.058
$p = 5\%$					
$\mathrm{N}(0, 100)$	1.133	1.302	1.582	1.042	3.178
Cauchy	0.939	1.155	1.559	1.038	4.119
$\mathrm{U}(-10, 10)$	1.008	1.105	1.522	0.991	3.420
$p = 10\%$					
$\mathrm{N}(0, 100)$	1.484	1.661	1.770	1.170	2.962
Cauchy	2.291	3.108	1.590	1.076	2.700
$\mathrm{U}(-10, 10)$	1.196	1.292	1.632	1.072	3.908
$p = 40\%$					
$\mathrm{U}(-20, 20)$	4.469	4.070	2.832	1.888	6.160

2.4 Example: GDP of China

The example deals with the annual values of Chinese GDP in the period 1952 - 2014 from National Bureau of Statistics of China (see Table 2.10). Since the GDP of China has a characteristic exponential growth, we apply a logarithmic transformation to the data. We employ Algorithm 2.2.2 for a linear trend. We choose the technical coefficients $b = 2.2$ and $W = 10$ since we suppose that the change points can appear in periods of approximately ten years.

The points detected by the algorithm as change points (namely 1961, 1982, 1994 and 2002) should correspond to the significant economic changes (see also Figure 2.3). In 1958, Mao Tse-tung announced the Great Leap Forward. Our model indicates this event in 1961. The constant term of our model decreases and the slope does not change so much in this year. It is interesting that we get absolutely identical results for a wide range of values b. Until 1978, our model shows a low but stable growth. After 1978, when the crucial reforms of the Chinese economy began, the growth of GDP speeds up. However, these reforms were realized only in several economic zones along the coast. A more rapid growth started in the early 1980s when the reforms were introduced in further areas. The growth was still quite fast in the 1990s, but it was accompanied by a high rate of inflation. In the period 2003-2006, other reforms (e.g., the protection of private property) were approved. In 2006, the 11th Five-Year Economic Program was accepted which aimed at education, medical care, etc. These changes of the Chinese economy are visible in GDP and the algorithm reflects them too.

Table 2.10. (GDP of China) Natural logarithm of Chinese GDP in 100 billions of Chinese yuan estimated by the improved linear sign test algorithm ($b = 2.2$, $W = 10$).

Year	1952	1953	1954	1955	1956	1957	1958	1959	1960	1961
ln(GDP)	11.13	11.32	11.36	11.42	11.54	11.58	11.78	11.88	11.89	11.71
Estimate	11.13	11.23	11.33	11.44	11.54	11.65	11.75	11.85	11.96	11.66
Year	1962	1963	1964	1965	1966	1967	1968	1969	1970	1971
ln(GDP)	11.65	11.73	11.89	12.05	12.14	12.09	12.06	12.18	12.33	12.40
Estimate	11.73	11.80	11.87	11.94	12.00	12.07	12.14	12.20	12.27	12.34
Year	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981
ln(GDP)	12.44	12.52	12.54	12.62	12.60	12.68	12.81	12.92	13.03	13.10
Estimate	12.41	12.47	12.54	12.61	12.67	12.74	12.81	12.88	12.94	13.01
Year	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991
ln(GDP)	13.19	13.30	13.49	13.71	13.85	14.01	14.23	14.35	14.45	14.60
Estimate	13.18	13.35	13.51	13.67	13.84	14.00	14.16	14.33	14.49	14.65
Year	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
ln(GDP)	14.81	15.08	15.39	15.63	15.78	15.89	15.95	16.01	16.12	16.22
Estimate	14.82	14.98	15.53	15.63	15.72	15.82	15.92	16.02	16.12	16.22
Year	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
ln(GDP)	16.31	16.43	16.59	16.74	16.90	17.10	17.27	17.36	17.53	17.70
Estimate	16.27	16.42	16.58	16.73	16.89	17.04	17.20	17.36	17.51	17.67
Year	2012	2013	2014							
Estimate	17.79	17.89	17.97							
ln(GDP)	17.82	17.98	18.13							

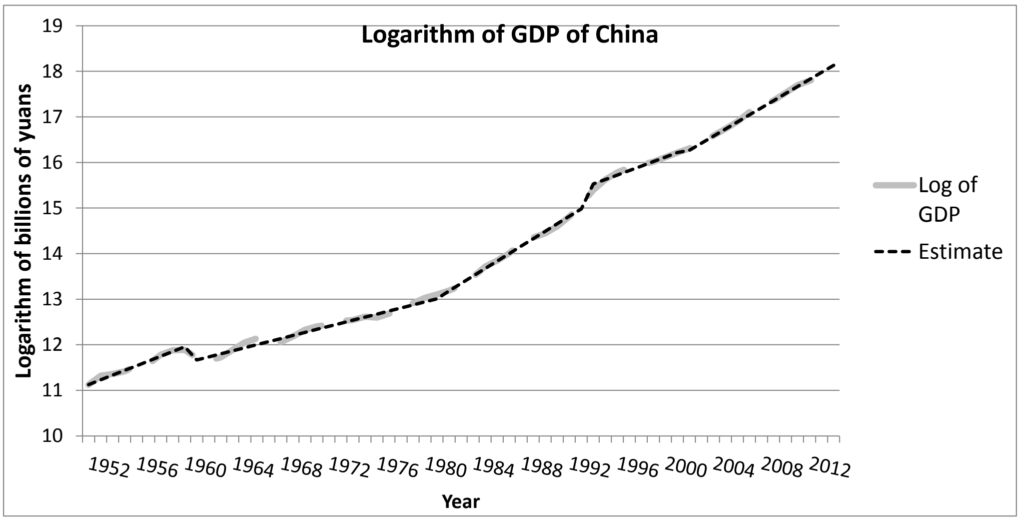


Figure 2.3. Natural logarithm of Chinese GDP in billions of Chinese yuan estimated by the improved linear sign test algorithm ($b = 2.2$, $W = 10$).

3. Generalization of trimmed mean based on geometric median

We propose new estimators of location in this chapter w. These estimators select a robust set around the geometric median also known as spatial median, enlarge it and compute the (iterative) weighted mean from it. By doing so, we obtain a robust estimator in the sense of the breakdown point which uses more observations than standard estimators. We apply our approach on the concepts of boxplot and bagplot. We work in a general normed vector space and allow multi-valued estimators.

The investigation of estimators from this chapter is also motivated by potential improvements of estimators from the previous chapter. We can employ e.g., in Algorithm 2.2.1 the estimators from this chapter instead of the simple median. However, this issue is also opened for future research.

We are interested in robust estimators for the parameter of location in this part. One of the first attempts to deal with such estimators are M-estimators, see (Huber, 1964) or (Maronna, 1976). They are computationally simple but suffer from low breakdown point, see (Hampel, 1971) or (Donoho and Huber, 1983). Its value is at most $\frac{1}{d+1}$, see (Maronna et al., 2006). Later, multiple estimators got proposed, among others we mention:

- minimum volume ellipsoid estimators (MVE, (Rousseeuw, 1985)) whose name stems from the fact that among all "proper" ellipsoids containing at least half of observations, the one given by MVE has minimal volume. However, their efficiency is rather poor.
- S-estimators ((Serfling, 1987)) have been suggested to overcome the low efficiency of MVE. They combine approaches of MVE and M-estimators.
- τ-estimators ((Lopuhaä, 1991)) also employ the idea of M-estimators but they do not require a preliminary scale estimator.
- Stahel-Donoho estimator ((Stahel, 1981) and (Donoho, 1982)) is based on the idea that any outlier in the multivariate case should be an outlier in some univariate projection.

The advantage of these estimators is their high breakdown point; the highest which a shift equivariant estimator can attain. However, their computation usually requires heavy effort. Therefore, (Maronna and Zamarb, 2002) suggested a way of reducing the computational complexity while sustaining the high breakdown point. As a price to pay, one is no longer able to estimate the covariance structure.

The early history of development and alternative approaches to spatial median are described in (Small, 1990). The work on spatial medians (generalisations of median) started on the Twelfth Census of the United States conducted in 1900. Statisticians of the time expressed an interest in studying the flow of population in the United States through the movement over time of a geographical center of the population. Maybe the earliest reference to the geometric median is to be found in (Hayford, 1902), where the vector of medians of orthogonal coordinates was suggested. However, the difficulty was recognized. This higher dimensional analog of the median is dependent on the choice of orthogonal coordinates used. (Scates, 1931) reexamined the problem of finding the geographical center of the United States, and found it (with the new concept of spatial median) to be '15 miles northwest of Dayton, Ohio'. Also other papers concerning this topic were published. However, they did not receive widespread attention. Therefore, it was rediscovered e.g., in (Haldane, 1948).

Many statistical estimators evolve from the method, where all observations are considered with the same weight, over the method where some of the observations are neglected, to the method where weights of the observations may differ. The evolution from the least squares over the least trimmed squares to the least weighted squares can serve as an example. See (Víšek, 2000) or (Víšek, 2011).

We will proceed in a similar fashion. We deal with the geometric median in the beginning, we employ further binary weights equal to 0 or 1 and in the end the weights from interval $[0, 1]$.

This chapter is organized as follows: we define the basic concept of geometric median in the first part of Section 3.1. Even though geometric median may be a set and not a point in general, most authors do not handle this fact. Because of this, we have decided to work with estimators which are multifunctions (also known as set-valued maps). The second part of Section 3.1 contains new results. We propose new estimators, discuss their breakdown point and provide a comparison between our algorithms and M-estimators.

We employ a slightly different notation in this chapter: often we will use the bold notation for $\boldsymbol{x} = (x_1, \dots, x_n) \in X^n$. We understand a component of a vector by the lower index while by the upper index, we mean an iteration number. A multifunction $R : X \rightrightarrows Y$ is a generalization of a function, where $R(x)$ does not have to be one point but may be a subset of Y. We say that R is bounded on bounded sets if $\cup_{x \in A} R(x)$ is a bounded set for all bounded sets $A \subset X$. Since we consider multi-valued estimators, some of the definitions are slightly generalized.

3.1 New estimators based on a generalization of trimmed mean

We first recall the geometric median in this section and on its basis derive other estimators. The basic idea is to find first the geometric median, then restrict ourselves to a set of neighboring observations and construct an estimator based only on this restricted set. If this set is chosen in a proper way, the estimator will have a breakdown point of $\frac{1}{2}$.

We continue with the definition of geometric median.

Definition 3.1.1. *We define the geometric median as a multifunction $\hat{T}_n : X^n \rightrightarrows X$ satisfying*

$$\hat{T}_n(x_1, \dots, x_n) = \operatorname*{argmin}_{a \in X} \sum_{j=1}^{n} \|a - x_j\|. \tag{3.1}$$

We present now two examples. The first one shows that the choice of the norm can change the geometric median in a significant way and that the geometric median may indeed be multi-valued. The second one depicts a simple situation where we are able to compute the geometric median. Moreover, it will be used later in some proofs.

Example 3.1.2. *Consider $X = \mathbb{R}^2$ and points $x_1 = (1, 0)$, $x_2 = (-1, 0)$ and $x_3 = x_4 = (0, 1)$. Then it is not difficult to verify that for the following norms we have*

$$\begin{aligned}(\mathbb{R}^2, \|\cdot\|_1) &\implies \hat{T}_4(x_1, \dots, x_4) = \operatorname{conv}\{(0,0), (0,1)\},\\ (\mathbb{R}^2, \|\cdot\|_2) &\implies \hat{T}_4(x_1, \dots, x_4) = \{(0,1)\},\\ (\mathbb{R}^2, \|\cdot\|_\infty) &\implies \hat{T}_4(x_1, \dots, x_4) = \{(0,1)\},\end{aligned}$$

where conv *stands for the convex hull. We see that for $\|\cdot\|_2$ and $\|\cdot\|_\infty$ the geometric median is determined in a unique way. This does not hold any more for $\|\cdot\|_1$.* $\triangle$

Example 3.1.3. *Consider $\bar{x} \in X$ and $\boldsymbol{x} = (x_1, \dots, x_n)$, where $x_1 = \cdots = x_m = \bar{x}$ for some $m \geq \frac{n}{2}$. Fix any $y \in X$. Then we have*

$$\begin{aligned}\sum_{i=1}^{n} \|\bar{x} - x_i\| = \sum_{i=m+1}^{n} \|\bar{x} - x_i\| &\leq \sum_{i=m+1}^{n} \|\bar{x} - y\| + \sum_{i=m+1}^{n} \|y - x_i\|\\ &= \sum_{i=m+1}^{n} \|\bar{x} - y\| + \sum_{i=m+1}^{n} \|y - x_i\| + \sum_{i=1}^{m} \|y - x_i\| - \sum_{i=1}^{m} \|y - \bar{x}\|\\ &= \sum_{i=1}^{n} \|y - x_i\| + (n - 2m)\|y - \bar{x}\| \leq \sum_{i=1}^{n} \|y - x_i\|\end{aligned}$$

due to the $m \geq \frac{n}{2}$. But this means that $\bar{x} \in \hat{T}_n(\boldsymbol{x})$. $\triangle$

In the next remark, we will mention connections between geometric median and other classical finite-dimensional concepts.

Remark 3.1.4. *The geometric median is a natural generalization of the median. Indeed, for $X = \mathbb{R}$ the geometric median is either a point or an interval. If it is a point, then it is the median. If it is an interval, then its midpoint equals to the median.*

For $X = \mathbb{R}^p$, there is also a close connection with the so-called depth, which measures "outlyingness" of a given multivariate sample, see (Liu, 1990). The type of depth was studied in (Zuo and Serfling, 2000), which can be expressed in the form

$$D_\Sigma(x; F) = \frac{1}{1 + \frac{1}{n}\sum_{i=1}^{n} \|x - x_i\|_{\Sigma^{-1}}},$$

where Σ is an covariance matrix of a distribution F and $\|x\|_{\Sigma^{-1}} := \sqrt{x^\top \Sigma^{-1} x}$ is a seminorm. It is clear that looking for a point maximizing $D_\Sigma(\cdot, F)$ is equivalent to searching for geometric median with respect to the corresponding seminorm. $\triangle$

Recall that a shift equivariant estimator T_n satisfies $T_n(x_1 + y, \dots, x_n + y) = T_n(x_1, \dots, x_n) + y$ for all $x_1, \dots, x_n \in X$ and $y \in X$. For single-valued estimators, it has been shown in (Maronna et al., 2006), formula (3.25) that a shift equivariant estimator satisfies

$$\varepsilon_n^*(T_n, \boldsymbol{x}) \leq \frac{1}{n}\left\lfloor \frac{n-1}{2} \right\rfloor. \tag{3.2}$$

Since the geometric median possesses this property, it is not surprising, that we obtain formula (3.2) as well. Moreover, we obtain even equality in this estimate.

Lemma 3.1.5. *For any $\boldsymbol{x} = (x_1, \dots, x_n) \in X^n$, we have for the geometric median*

$$\varepsilon_n^*(\hat{T}_n, \boldsymbol{x}) = \frac{1}{n}\left\lfloor \frac{n-1}{2} \right\rfloor,$$

and thus for the asymptotic breakdown point we have $\varepsilon^* = \frac{1}{2}$.

Proof. The second statement is an immediate consequence of the first one. Note that the first statement is equivalent to $m_n^*(\hat{T}_n, \boldsymbol{x}) = n_0$ with $n_0 := \lfloor \frac{n-1}{2} \rfloor$. We see from Example 3.1.3 that $m_n^*(\hat{T}_n, \boldsymbol{x}) < \frac{n}{2}$, which further implies $m_n^*(\hat{T}_n, \boldsymbol{x}) \le \frac{n}{2} - 1 \le n_0$. To finish the proof, it is sufficient to show that $m_n^*(\hat{T}_n, \boldsymbol{x}) \ge n_0$.

Consider thus any $\tilde{\boldsymbol{x}} \in A_{n_0,n}(\boldsymbol{x})$ and denote by I the index set of coordinates where $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ differ and by J its complement. Denote by n_1 the cardinality of I and observe that $n_1 \le n_0$. Denoting further $R := \max_{i=1,\dots,n} \|x_i\|$, we have $\|\tilde{x}_j\| = \|x_j\| \le R$ for all $j \in J$. Taking any $j \in J$ and $y \in X$, we obtain the following estimate

$$\begin{aligned}
\sum_{l=1}^n \|\tilde{x}_j - \tilde{x}_l\| \le \sum_{l\in I} \|\tilde{x}_j - \tilde{x}_l\| + 2(n-n_1)R &\le \sum_{l\in I} \|\tilde{x}_j - y\| + \sum_{l\in I} \|y - \tilde{x}_l\| + \\
&+ 2(n-n_1)R \\
&= \sum_{l=1}^n \|y - \tilde{x}_l\| - \sum_{l\in J} \|y - \tilde{x}_l\| + n_1\|\tilde{x}_j - y\| + 2(n-n_1)R \\
&\le \sum_{l=1}^n \|y - \tilde{x}_l\| - \sum_{l\in J} \|y - \tilde{x}_j\| + \sum_{l\in J} \|\tilde{x}_j - \tilde{x}_l\| + n_1\|\tilde{x}_j - y\| + \\
&+ 2(n-n_1)R \\
&\le \sum_{l=1}^n \|y - \tilde{x}_l\| + (2n_1 - n)\|y - \tilde{x}_j\| + 4(n-n_1)R
\end{aligned}$$

Since $2n_1 - n \le 2n_0 - n < 0$, we obtain that there is $\tilde{R}_I > 0$ such that for all $\|y\| \ge \tilde{R}_I$ we have

$$\sum_{l=1}^n \|\tilde{x}_j - \tilde{x}_l\| < \sum_{l=1}^n \|y - \tilde{x}_l\|$$

But this means that the geometric median lies in a ball with radius $\tilde{R}_I$. Since there is only a finite number of possible subsets I, we have finished the proof. □

The next lemma allows us to compute the breakdown point of an estimator.

Lemma 3.1.6. *Consider multifunctions $\Phi_1 : X^n \rightrightarrows X^m$ and $\Phi_2 : X^n \times X^m \rightrightarrows X$. Assume that the following assumptions are satisfied:*

1. *All components of Φ_1 have breakdown point at least p.*
2. *There exists $\Phi_3 : X^m \rightrightarrows X$ which is bounded on bounded sets such that $\|\Phi_2(\boldsymbol{x}, \boldsymbol{y})\| \le \|\Phi_3(\boldsymbol{y})\|$ for all $\boldsymbol{x} \in X^n$ and $\boldsymbol{y} \in X^m$.*

Then estimator T_n defined as

$$T_n(\boldsymbol{x}) := \bigcup_{\boldsymbol{y}\in\Phi_1(\boldsymbol{x})} \Phi_2(\boldsymbol{x},\boldsymbol{y})$$

has a breakdown point of at least p.

Proof. Due to the first assumption, there exists some $R > 0$ such that for $m := np$, all $\tilde{\boldsymbol{x}} \in A_{m,n}(\boldsymbol{x})$ and for all $\boldsymbol{y} \in \Phi_1(\tilde{\boldsymbol{x}})$ we have $\|\boldsymbol{y}\| \le R$. But then we have $\|\Phi_2(\tilde{\boldsymbol{x}},\boldsymbol{y})\| \le \|\Phi_3(\boldsymbol{y})\|$, which is uniformly bounded due to the second assumption. Thus, the statement has been proved. □

We come now to new estimators. For a set $S \subset X$ and a point $\boldsymbol{x} = (x_1,\dots,x_n) \in X^n$ we define

$$\mathcal{L}(S,\boldsymbol{x}) :=$$
$$\bigcup_{y\in S}\left\{\boldsymbol{x}_I \in X^{\left\lfloor\frac{n-1}{2}\right\rfloor} \,\middle|\, \exists I \subset \{1,\dots,n\} :\ \max_{i\in I}\|x_i - y\| \le \min_{i\in\{1,\dots,n\}\setminus I}\|x_i - y\|\right\},$$

where $\boldsymbol{x}_I$ denotes the restriction of $\boldsymbol{x}$ to components I. The interpretation of this set goes as follows: we select some $y \in S$ and choose $\boldsymbol{x}_I$ to be the $\left\lfloor\frac{n-1}{2}\right\rfloor$ observations closest to y. Then $\mathcal{L}(S,\boldsymbol{x})$ is the union of all such subsets with respect to all choices of $y \in S$. Since every such $\boldsymbol{x}_I$ contains less than $\frac{n}{2}$ components of $\boldsymbol{x}$, this set is stable with respect to perturbations of $\boldsymbol{x}$ whenever less than one half of the observations are contaminated.

We will use $\mathcal{L}(S,\boldsymbol{x})$ to define further estimators. The next theorem says that if we start with the geometric median $S = \hat{T}_n(\boldsymbol{x})$ and a multifunction R with certain boundedness properties, we obtain an estimator with the same breakdown point as the geometric median.

Theorem 3.1.7. *Consider any multifunction $R : X^{\left\lfloor\frac{n-1}{2}\right\rfloor} \rightrightarrows X$ which is bounded on bounded sets and for which there exists z^k such that $\|R(z^k,\dots,z^k)\| \to \infty$. Then for estimator $T_n : X^n \rightrightarrows X$ defined as*

$$T_n^1(\boldsymbol{x}) := \bigcup_{\boldsymbol{y}\in\mathcal{L}(\hat{T}_n(\boldsymbol{x}),\boldsymbol{x})} R(\boldsymbol{y}) \tag{3.3}$$

and for every $\boldsymbol{x} = (x_1,\dots,x_n) \in X^n$ we have the following relation

$$\varepsilon_n^*(T_n^1,\boldsymbol{x}) = \varepsilon_n^*(\hat{T}_n,\boldsymbol{x}) = \frac{1}{n}\left\lfloor\frac{n-1}{2}\right\rfloor.$$

Proof. We obtain

$$\varepsilon_n^*(T_n^1,\boldsymbol{x}) \ge \varepsilon_n^*(\hat{T}_n,\boldsymbol{x}) = \frac{1}{n}\left\lfloor\frac{n-1}{2}\right\rfloor$$

from Lemma 3.1.6 with $m = \left\lfloor\frac{n-1}{2}\right\rfloor$, $\Phi_1(\boldsymbol{x}) = \mathcal{L}(\hat{T}_n(\boldsymbol{x}),\boldsymbol{x})$ and $\Phi_2(\boldsymbol{x},\boldsymbol{y}) = R(\boldsymbol{y})$ and Lemma 3.1.5. To show the opposite inequality, realize that the statement is equivalent to $m_n^*(T_n^1,\boldsymbol{x}) \le \left\lfloor\frac{n-1}{2}\right\rfloor$. Assume for contradiction that $m_n^*(T_n^1,\boldsymbol{x}) \ge \left\lfloor\frac{n-1}{2}\right\rfloor + 1 \ge \left\lfloor\frac{n}{2}\right\rfloor$. We change the first $\left\lfloor\frac{n}{2}\right\rfloor$ coordinates of $\boldsymbol{x}$ to z^k and denote the perturbed point by $\tilde{\boldsymbol{x}}^k$. Then Example 3.1.3 tells us that $z^k \in \hat{T}_n(\tilde{\boldsymbol{x}}^k)$. Due to the definition of $\mathcal{L}$, we see that $(z^k,\dots,z^k) \in \mathcal{L}(\hat{T}_n(\tilde{\boldsymbol{x}}^k),\tilde{\boldsymbol{x}}^k)$ and the imposed assumption of R implies a contradiction. □

If both R and $\mathcal{L}(\hat{T}_n(\boldsymbol{x}), \boldsymbol{x})$ are single-valued functions, expression (3.3) reduces to

$$T_n^1(\boldsymbol{x}) = R(\mathcal{L}(\hat{T}_n(\boldsymbol{x}), \boldsymbol{x})).$$

Moreover, in such a case $\mathcal{L}(\hat{T}_n(\boldsymbol{x}), \boldsymbol{x})$ denotes one half of observations which are closest to the geometric median. There are several natural choices for R: for example mean, weighted mean or geometric median.

One of the possible drawbacks of estimator (3.3) is that it utilizes only half of the original data. We want to make use of as many observations as possible while maintaining the high breakdown point. To this aim, we first consider a general set $\mathcal{S} \subset X^m$ for some $m \in \mathbb{N}$, for example we may consider mean of $\boldsymbol{x}$ as a subset of X or $\mathcal{L}(\hat{T}_n(\boldsymbol{x}), \boldsymbol{x})$ as a subset of $X^{\left\lfloor \frac{n-1}{2} \right\rfloor}$. Then we consider some $b : X \times X^n \to [0, \infty)$ and enlarge $\mathcal{S}$ by defining

$$\mathcal{E}_b(\mathcal{S}, \boldsymbol{x}) := \bigcup_{\boldsymbol{y} \in \mathcal{S}} \left\{ \boldsymbol{x}_I | \; I = \{ i | \; x_i \in \cup_{j=1}^m \mathbb{B}(y_j, b(y_j, \boldsymbol{x})) \} \right\}. \tag{3.4}$$

Here, $\mathbb{B}(y_j, b(y_j, \boldsymbol{x}))$ stands for a ball around y_j with radius $b(y_j, \boldsymbol{x})$. The interpretation goes as follows: from $\mathcal{S}$ we select $\boldsymbol{y}$, make balls around all of its components and select all components of $\boldsymbol{x}$ which lie in the union of these balls.

Example 3.1.8. *Consider the case of $X = \mathbb{R}$, $n = 5$ and $\boldsymbol{x} = (-3, -2, 0, 2, 4)$. Then the geometric median equals to $\hat{T}_n(\boldsymbol{x}) = 0$ and since $n_0 = \left\lfloor \frac{n-1}{2} \right\rfloor = 2$, we also have*

$$\mathcal{L}(\hat{T}_n(\boldsymbol{x}), \boldsymbol{x}) = \{(-2, 0), (0, 2)\} \subset \mathbb{R}^2.$$

If we consider $b \equiv 1$, then

$$\mathcal{E}_b(\mathcal{L}(\hat{T}_n(\boldsymbol{x}), \boldsymbol{x}), \boldsymbol{x}) = \{(-3, -2, 0), (0, 2)\}.$$

Note that both elements of $\mathcal{E}_b(\mathcal{L}(\hat{T}_n(\boldsymbol{x}), \boldsymbol{x}), \boldsymbol{x})$ are of a different dimension. △

We obtain the following variant of Theorem 3.1.7, for which we omit its identical proof.

Theorem 3.1.9. *Consider $b : X \times X^n \to [0, \infty)$ bounded on bounded sets in the first variable, uniformly in the second one, any family of multifunctions $R_s : X^s \rightrightarrows X$ for $s = 1, \ldots, n$ which are all bounded on bounded sets and for which there exists z^k such that $\|R_s(z^k, \ldots, z^k)\| \to \infty$. Then for estimators $T_n^2 : X^n \rightrightarrows X$ and $T_n^3 : X^n \rightrightarrows X$ defined as*

$$T_n^2(\boldsymbol{x}) := \bigcup_{\boldsymbol{y} \in \mathcal{E}_b(\hat{T}_n(\boldsymbol{x}), \boldsymbol{x})} R_{\dim \boldsymbol{y}}(\boldsymbol{y}), \tag{3.5a}$$

$$T_n^3(\boldsymbol{x}) := \bigcup_{\boldsymbol{y} \in \mathcal{E}_b(\mathcal{L}(\hat{T}_n(\boldsymbol{x}), \boldsymbol{x}), \boldsymbol{x})} R_{\dim \boldsymbol{y}}(\boldsymbol{y}) \tag{3.5b}$$

and for every $\boldsymbol{x} = (x_1, \ldots, x_n) \in X^n$, we have the following relation for breakdown points

$$\varepsilon_n^*(T_n^2, \boldsymbol{x}) = \varepsilon_n^*(T_n^3, \boldsymbol{x}) = \varepsilon_n^*(\hat{T}_n, \boldsymbol{x}) = \frac{1}{n} \left\lfloor \frac{n-1}{2} \right\rfloor.$$

Function b should neither have too large values (which corresponds to a vast enlargement of the set in question) because outliers may be close to the non-contaminated data, nor too small values because some information could be missed. We suggest a possible choice in the Appendix.

The crucial question lies in the choice of R. There are several natural possibilities, we can use mean, weighted mean with given weights w or geometric median. This leads to the following estimators:

$$T_n^1(\boldsymbol{x}) := \bigcup_{\boldsymbol{y}\in\mathcal{E}_b(\mathcal{L}(\hat{T}_n(\boldsymbol{x}),\boldsymbol{x}),\boldsymbol{x})} \left\{\frac{1}{n_0}\sum_{i=1}^{n_0} y_i\right\}, \tag{3.6a}$$

$$T_n^2(\boldsymbol{x}) := \bigcup_{\boldsymbol{y}\in\mathcal{E}_b(\mathcal{L}(\hat{T}_n(\boldsymbol{x}),\boldsymbol{x}),\boldsymbol{x})} \left\{\sum_{i=1}^{n_0} w_i y_i\right\}, \tag{3.6b}$$

$$T_n^3(\boldsymbol{x}) := \bigcup_{\boldsymbol{y}\in\mathcal{E}_b(\mathcal{L}(\hat{T}_n(\boldsymbol{x}),\boldsymbol{x}),\boldsymbol{x})} \hat{T}_{n_0}(\boldsymbol{y}). \tag{3.6c}$$

Similarly to the geometric median, we obtain from Theorem 3.1.9 that all the estimates in (3.6) have the breakdown point of $\frac{1}{n}\left\lfloor\frac{n-1}{2}\right\rfloor$ and the limiting breakdown point of $\frac{1}{2}$.

To improve the behavior of the estimators, we implement an iterative procedure. We start with geometric median $z^0 = \hat{T}_n(\boldsymbol{x})$ and in every iteration k compute a new estimate z^k. To do so, we employ (3.5b) with R being the weighted mean, where the (non-normalized) weights satisfy

$$w_i(z^{k-1}, y_i) = \begin{cases} 1 & \text{if } y_i \in \mathcal{L}(\{z^{k-1}\},\boldsymbol{x}), \\ \max\limits_{y_j\in\mathcal{L}(\{z^{k-1}\},\boldsymbol{x})} \left(1 - \frac{\|y_i-y_j\|}{b^k(y_j,\boldsymbol{x})}\right) & \text{otherwise} \end{cases} \tag{3.7}$$

for some b^k based on $\mathcal{L}(\{z^{k-1}\},\boldsymbol{x})$. This choice of weights makes use of the possibly division of components of $\boldsymbol{y}\in\mathcal{E}_b(\mathcal{L}(\{z^{k-1}\},\boldsymbol{x}),\boldsymbol{x})$ into two parts: those who belong to $\mathcal{L}(\{z^{k-1}\},\boldsymbol{x})$ and those who were added by enlarging this set. We choose the (non-normalized) weight equal to one for the first part, the weight for observations from the second part decreases with the increasing distance to $\mathcal{L}(\{z^{k-1}\},\boldsymbol{x})$. We summarize this approach in Algorithm 3.1.1. Considering the termination criterion, any standard criterion may be used, for example if the (relative) change in z^k is small.

Finally, we would like to point out that every update step in Algorithm 3.1.1 keeps the stability result which we already mentioned several times in the previous text. For $k=1$, we may write

$$z^k = \bigcup_{\boldsymbol{y}\in\Phi_1(\boldsymbol{x})} \Phi_2(\boldsymbol{x},\boldsymbol{y}),$$

where $\boldsymbol{y}=\boldsymbol{y}^1$, $\Phi_1(\boldsymbol{x}) := \mathcal{E}_b(\mathcal{L}(\{z^0\},\boldsymbol{x}),\boldsymbol{x})$ and $\Phi_2(\boldsymbol{x},\boldsymbol{y}) := \sum_{i=1}^{\dim \boldsymbol{y}^1} w_i(\boldsymbol{x},\boldsymbol{y}^1)y_i^1$. Then Φ_1 has breakdown point $\frac{1}{n}\left\lfloor\frac{n-1}{2}\right\rfloor$ and since $\|\Phi_2(\boldsymbol{x},\boldsymbol{y})\| \le n\max_{y_i\in\boldsymbol{y}}\|y_i\|$ holds true, thanks to Lemma 3.1.6 we obtain that z^1 has the same breakdown point. By applying the same procedure to subsequent iterations, we obtain the same result for all z^k.

Algorithm 3.1.1 An estimator based on iterative weighting

```
Input: observations $\boldsymbol{x} = (x_1, \dots, x_n)$
$k \leftarrow 0$, $z^0 \leftarrow \hat{T}_n(\boldsymbol{x})$
while not terminate do
    $k \leftarrow k + 1$
    determine $b^k$ based on $\mathcal{L}(\{z^{k-1}\}, \boldsymbol{x})$
    pick any $\boldsymbol{y}^k \in \mathcal{E}_{b^k}(\mathcal{L}(\{z^{k-1}\}, \boldsymbol{x}), \boldsymbol{x})$
    compute $w^k$ according to formula (3.7) and renorm them such that their sum equals to 1
    $z^k \leftarrow \sum_{i=1}^{\dim \boldsymbol{y}} w_i^k y_i^k$
end while
return estimate $\hat{x} \leftarrow z^k$
```

In the previous text we highlighted some benefits of our estimators. Note that naturally there are also situations where it is better to use standard estimators. Consider for example the one-dimensional double exponential distribution with density $\frac{1}{2}\exp(-|x - \mu|)$. Then the (geometric) median coincides with the maximum likelihood estimator of μ, see Section 6.3 in (Lehmann and Casella, 2006), and therefore the median is the most efficient estimator. Thus, by employing additional observations apart from the median we only worsen the quality of an estimator. However, to benefit from such situation, we would have to know the true distribution and know that there is no contamination. We illustrate this in Table 3.4, where it is visible that for some heavy tailed distributions median outperforms other estimates.

We deal with equivariance of our estimators in the following lemma.

Lemma 3.1.10. *Suppose that the assumptions of Theorem 3.1.7 are satisfied and that R is shift and scale equivariant function, then $T_n^1(\boldsymbol{x})$ defined in* (3.3) *is shift and scale equivariant estimator.*

Proof. Geometric median is shift and scale equivariant, therefore the set $\mathcal{L}(\hat{T}_n(\boldsymbol{x}), \boldsymbol{x})$ changes in accordance. It suffices now to consider equivariance of the function R. □

For its simplicity, we omit the proof of the following lemma.

Lemma 3.1.11. *Suppose that the assumptions of Theorem 3.1.9 are satisfied and that R is shift and scale equivariant function, then $T_n^2(\boldsymbol{x})$ and $T_n^2(\boldsymbol{x})$ defined in* (3.5) *are shift and scale equivariant estimators.*

Remark 3.1.12. *There is some connection between our estimators with M-estimators. We summarize the algorithm from (Maronna et al., 2006) in Algorithm 3.1.2, Section 2.7.3.*

Both approaches (iteratively) compute a weighted mean of observations. While M-estimators are based on the maximum likelihood estimate, ours are based on geometric intuition and trimmed mean. If $X = \mathbb{R}$ and if W_1 in Algorithm 3.1.2 has finite support, we can say that our estimators belong to the very wide class of M-estimators. This changes for $\mathbb{R}^d$ though. Under the standard assumption that W_1 is symmetric around zero and non-increasing on rays emanating from zero, the weight of an observation depends only on the distance from z^{k-1}. Thus,

Algorithm 3.1.2 M-estimator from (Maronna et al., 2006), Section 2.7.3

```
Input: observations $\boldsymbol{x} = (x_1, \ldots, x_n)$, weighting functions $W_1$ and $W_2$
$k \leftarrow 0$, $z^0 \leftarrow \hat{T}_n(\boldsymbol{x})$, dispersion estimate $\sigma^0$
while not terminate do
    $k \leftarrow k + 1$
    $r_i^k \leftarrow \frac{x_i - z^{k-1}}{\sigma^{k-1}}$
    $w_{1,i}^k \leftarrow W_1(r_i^k)$, $w_{2,i}^k \leftarrow W_2(r_i^k)$ and norm the weights to sum to one
    $z^k \leftarrow \sum_{i=1}^n w_{1,i}^k x_i$, $\sigma^k \leftarrow \frac{1}{n} \sum_{i=1}^n w_{2,i}^k (x_i - z^{k-1})^2$
end while
return estimate $\hat{x} \leftarrow z^k$
```

two observations have the same weight if and only if their distance to z^{k-1} is identical. On the other hand, in our approach all points in $\mathcal{L}(\{z^{k-1}\}, \boldsymbol{x})$ have the same weight and this set is enlarged farther for distant observations. Thus, even though none of the algorithms estimates the covariance structure, our algorithm makes at least an attempt to consider it.

Of course, there are M-estimators which along with the location also properly estimate the covariance structure. But this raises the computational complexity and reduces the breakdown point to $\frac{1}{d+1}$. To summarize: we can say that our algorithms try to pick the best properties of M-estimators, on the one hand they have high breakdown point and are simple to compute; on the other hand, they at least partially consider the covariance structure.

Another advantage of our estimator over M-estimators is a simpler theoretical analysis. To show that our estimator has the limiting breakdown point $\frac{1}{2}$, it is sufficient to apply Lemma 3.1.6, which itself directly follows from the definition of the breakdown point. To the best of our knowledge, such direct application of the definition is not possible for M-estimators, for example one has to take care of properties of weighting function due to the division by dispersion. △

3.2 Consistency

We want to study the consistency of our estimators in this part. See Definition 1.1.3. We suppose that the observations $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ are i.i.d. with the same distribution as a random variable Z. The theoretical properties of geometric median were studied in (Kemperman, 1987). The theoretical counterpart was employed there in the form

$$\hat{T}(Z) = \operatorname*{argmin}_{a \in X} \operatorname{E} \|a - Z\| - \|Z\|, \tag{3.8}$$

where Z is a random variable. The definition of theoretical geometric median from (3.8) has the advantage that we do not need a finiteness of the first moment.

The theoretical properties of geometric median were studied in (Kemperman, 1987) for the case of X to be separable Hilbert space. The uniqueness of $\hat{T}(Z)$ under the condition that the random variable Z is not concentrated on a straight line was shown there.

We can write

$$\operatorname*{argmin}_{a\in X} \sum_{j=1}^{n} \|a - x_j\| = \operatorname*{argmin}_{a\in X} \frac{1}{n} \sum_{j=1}^{n} \|a - x_j\|.$$

From that follows

$$\operatorname*{argmin}_{a\in X} \sum_{j=1}^{n} \|a - x_j\| \xrightarrow{\text{a.s.}} \operatorname*{argmin}_{a\in X} \mathrm{E}\,\|Z - a\|,$$

where $\xrightarrow{\text{a.s.}}$ denotes almost sure convergence. However, we have to suppose $\mathrm{E}\,\|Z - a\| < \infty$. If it holds then

$$\operatorname*{argmin}_{a\in X} \mathrm{E}\,\|Z - a\| = \operatorname*{argmin}_{a\in X} \mathrm{E}\,\|Z - a\| - \|Z\|.$$

Therefore, the definition from (Kemperman, 1987) is more suitable also in our case.

Lemma 3.2.1. *Let $\boldsymbol{z} = (z_1, z_2 \dots)$ be i.i.d. observations of the variable $Z : (\Omega, \mathcal{A}, \mathrm{P}) \to X$.*

1. *Let $T_n(\boldsymbol{z})$ be a consistent estimator of the parameter $T(Z)$ on the parameter space Υ,*
2. *$S_n(T(Z), \boldsymbol{z})$ be a consistent estimator of $S(T(Z), Z)$ from parameter space Θ for any $T(Z) \in \Upsilon$*
3. *and for any $\varepsilon > 0$, $\delta > 0$ and $U \in \Upsilon$ there is $\zeta > 0$ such that for all n and A in ζ neighborhood of U $\mathrm{P}(\|S_n(U, \boldsymbol{z}) - S_n(A, \boldsymbol{z})\| > \varepsilon) < \delta$,*

then $S_n(T_n(\boldsymbol{z}), \boldsymbol{z})$ is a consistent estimator of $S(T(Z), Z)$. For the sake of this lemma, we will not distinguish by notation between norms of Θ and Υ and we denote them as $\|\cdot\|$.

Proof. We want to show that for any $\varepsilon > 0$ and $\delta > 0$ there is $n_0 \in \mathbb{N}$ such that for all $n \in \mathbb{N}$, $n \geq n_0$ holds $\mathrm{P}(\|S_n(T_n(\boldsymbol{z}), \boldsymbol{z}) - S(T(Z), Z)\| > \varepsilon) < \delta$

$$\begin{gathered}\mathrm{P}(\|S_n(T_n(\boldsymbol{z}), \boldsymbol{z}) - S(T(Z), Z)\| > \varepsilon) \leq \\ \mathrm{P}(\|S_n(T_n(\boldsymbol{z}), \boldsymbol{z}) - S_n(T(Z), \boldsymbol{z})\| + \|S_n(T(Z), \boldsymbol{z}) - S(T(Z), Z)\| > \varepsilon),\end{gathered} \tag{3.9}$$

where we have employed triangle inequality. (3.9)

We define

$$\begin{aligned} E_1 &= \left\{\omega \,\middle|\, \|S_n(T_n(\boldsymbol{z}), \boldsymbol{z}) - S_n(T(Z), \boldsymbol{z})\| > \frac{\varepsilon}{2}\right\}, \\ E_2 &= \left\{\omega \,\middle|\, \|S_n(T(Z), \boldsymbol{z}) - S(T(Z), Z)\| > \frac{\varepsilon}{2}\right\}. \end{aligned}$$

The expression in (3.9) is lees than

$$\mathrm{P}(E_1 \cup E_2) \leq \mathrm{P}(E_1) + \mathrm{P}(E_2).$$

We exploit now Assumption 3 of our lemma to find $\zeta > 0$ such that for all n and A from ζ neighborhood of $T(Z)$ we have $\mathrm{P}(\|S_n(A,\boldsymbol{z}) - S_n(T(Z),\boldsymbol{z})\| > \frac{\varepsilon}{2}) < \frac{\delta}{3}$.

We find n_0 such that for all $n \geq n_0$ we get $\mathrm{P}(\|T_n(\boldsymbol{z}) - T(Z)\| > \zeta) < \frac{\delta}{3}$. This can be done according to Assumption 1.

For $n \geq n_0$

$$\begin{aligned} \mathrm{P}(E_1) &\leq \mathrm{P}(\|T_n(\boldsymbol{z}) - T(Z)\| > \zeta) + \\ &+ \mathrm{P}(\|S_n(T_n(\boldsymbol{z}),\boldsymbol{z}) - S_n(T(Z),\boldsymbol{z})\| > \frac{\varepsilon}{2}, \|T_n(\boldsymbol{z}) - T(Z)\| \leq \zeta) \leq \frac{2\delta}{3}. \end{aligned}$$

The term $\mathrm{P}(E_2)$ is smaller than $\frac{\delta}{3}$ thanks to 2. □

We deal now with theoretical counterpart of $\boldsymbol{\mathcal{L}}$. The following proposition asserts that we can define theoretical counterpart of $\boldsymbol{y} \in \boldsymbol{\mathcal{L}}$ as

$$\left\{ Z(\omega) \middle| \|Z(\omega) - \hat{T}(Z)\| \leq a \right\}$$

for some appropriately chosen constant a.

Proposition 3.2.2. *Consider continuous random variable $Z : (\Omega, \mathcal{A}, \mathrm{P}) \to \mathbb{R}^d$. We denote its geometric median as $\hat{T}(Z)$. We have random sample $\boldsymbol{z} = (z_1, z_2, \dots)$ of observations from Z. $d = \max_{y \in \boldsymbol{y}} \|y - \hat{T}_n(\boldsymbol{z})\|$ is then for any $\boldsymbol{y} \in \boldsymbol{\mathcal{L}}(\hat{T}_n(\boldsymbol{z}), \boldsymbol{z})$ a consistent estimator of*

$$a = \min_{b>0} \left(\mathrm{P}(\|Z - \hat{T}(Z)\| \leq b) = \frac{1}{2} \right). \tag{3.10}$$

Proof. Since Z is the continuous random variable, $\|Z - \hat{T}(Z)\|$ is also the continuous random variable, therefore $\min_{b>0} \left(\mathrm{P}(\|Z(\omega) - \hat{T}(Z)\| \leq b) = \frac{1}{2} \right)$ exists and a is well defined.

We will prove three assumptions from Lemma 3.2.1.

The consistence of geometric median was discussed before.

We will prove now Assumption 2 of the lemma. Consider any $\varepsilon > 0$. Define $\phi = \mathrm{P}(\|z_i - \hat{T}(Z)\| > a + \varepsilon)$ and V_n as the number of $\|z_i - \hat{T}(Z)\|$ which are greater than $a + \varepsilon$ $(i = 1, \dots, n)$. From the definition of a follows that $\phi < \frac{1}{2}$ and that V_n has binomial distribution with parameters ϕ and n. Let us consider n

odd.

$$\begin{aligned}\mathrm{P}(\max_{z_i\in\boldsymbol{y}}\|z_i-\hat{T}(Z)\|\geq a+\varepsilon) &= \mathrm{P}(V_n\geq\frac{n+1}{2})\\ &= \mathrm{P}(V_n-n\phi\geq\frac{n+1}{2}-n\phi)\\ &= \mathrm{P}(V_n-n\phi\geq n(\frac{1}{2}-\phi)+\frac{1}{2})\\ &\leq \mathrm{P}(V_n-n\phi\geq n(\frac{1}{2}-\phi))\\ &\leq \mathrm{P}(|V_n-n\phi|\geq n(\frac{1}{2}-\phi))\\ &= \mathrm{P}((V_n-n\phi)^2\geq n^2(\frac{1}{2}-\phi)^2)\\ &\leq \frac{\mathrm{E}(V_n-n\phi)^2}{n^2(\frac{1}{2}-\phi)^2}\\ &= \frac{n\phi(1-\phi)}{n^2(\frac{1}{2}-\phi)^2}.\end{aligned}$$

We employ Markov's inequality in the last inequality.

Since $\phi<\frac{1}{2}$, we see that the last term on the right hand side converges to zero as $n\to\infty$. The case when n is even can be proved in a similar way.

We have to deal also with the case $\mathrm{P}(\min_{z_i\notin\boldsymbol{y}}\|z_i-\hat{T}(Z)\|\leq a-\varepsilon)$. We define similarly as before F_n as the number of observations z_i for which $\|z_i-\hat{T}(Z)\|\leq a-\varepsilon$ and $i=1,\ldots,n$. Further, we define $\varphi=\mathrm{P}(\|z_i-\hat{T}(Z)\|\leq a-\varepsilon)$. We get $\varphi<\frac{1}{2}$ from the definition of a. F_n has the binomial distribution with parameters n and φ. We can put

$$\mathrm{P}(\min_{z_i\notin\boldsymbol{y}}\|z_i-\hat{T}(Z)\|\leq a-\varepsilon)=\mathrm{P}(F_n\geq\frac{n+1}{2})$$

and continue in the same way as in the previous case.

We have to deal with Assumption 3 of the lemma now. But we get for $U,A\in\mathbb{R}^p$

$$\max_{y\in\boldsymbol{y}}\|y-U\|\leq\max_{y\in\boldsymbol{y}}(\|y-A\|+\|A-U\|)=\max_{y\in\boldsymbol{y}}(\|y-A\|)+\|A-U\|.$$

It yields $|\max_{y\in\boldsymbol{y}}\|y-U\|-\max_{y\in\boldsymbol{y}}\|y-A\||\leq\|A-U\|$. Thus, we can find ζ (by the choice of U and A) such that $|\max_{y\in\boldsymbol{y}}\|y-U\|-\max_{y\in\boldsymbol{y}}\|y-A\||$ is less than ζ for all n. □

We replace the assumption about continuity of Z for general $Z:(\Omega,\mathcal{A},\mathrm{P})\to(X,\|\cdot\|)$ by the assumption that $\mathrm{P}(\|Z\|\leq b)$ is continuous with respect to b. However, this still does not take into account discrete functions etc.

The previous proposition says that $y\in\boldsymbol{y}$ $\|y-\hat{T}_n(\boldsymbol{z})\|\leq a$ holds for large n and $\boldsymbol{y}\in\mathcal{L}(\hat{T}_n(\boldsymbol{z}),\boldsymbol{z})$.

Remark 3.2.3. *Let us consider $\mathcal{E}_b(\mathcal{S},\boldsymbol{x})$ as in Equation (3.4) and Z continuous. We denote $f(x)$ as the density function of Z, where $x\in X$. $\mathcal{E}_b(\mathcal{S},\boldsymbol{x})\subset\bigcup_{\boldsymbol{y}\in\mathcal{S}}\left\{x|\ x\in X,f(x)>0,x\in\cup_{j=1}^m\mathbb{B}(y_j,b(y_j,\boldsymbol{x}))\right\}$ for large n . Because the set $\bigcup_{\boldsymbol{y}\in\mathcal{S}}\left\{x|\ x\in X,f(x)>0,x\in\cup_{j=1}^m\mathbb{B}(y_j,b(y_j,\boldsymbol{x}))\right\}$ can be uncountable, the equality does not have to be fulfilled by $\mathcal{E}_b(\mathcal{S},\boldsymbol{x})$.*

Similar consideration can be done for $\mathcal{L}$.

3.3 Numerical results

We first show the numerical performance of our estimators in this section and then how they comply with the well-known concepts of boxplot and bagplot.

3.3.1 The numerical performance

We consider $X = \mathbb{R}^d$ with $d \in \{1, 15\}$ and compare our algorithms with known estimators; for reader's convenience we summarize the used algorithms in Table 3.1.

Table 3.1. Summary of algorithms. The horizontal line divides known algorithms from our own.

Mean	mean
Med	median or geometric median
Trun	α-truncated mean with $\alpha = 0.2$
Winsor	α-winsorized mean with $\alpha = 0.2$
M1	Huber M-estimator, see (Huber, 1981)
M2	Algorithm 3.1.2 from (Maronna et al., 2006)
SD	Stahel-Donoho estimator, see (Stahel, 1981) and (Donoho, 1982)
GM1	formula (3.3), where R is the mean
GM2	formula (3.5a), where R is the mean
GM3	Algorithm 3.1.1

To generate the samples, we first generate z_i from $\mathrm{N}(0,1)$, then contaminate them by some distribution with probability p and finally modify them via a covariance structure. This modification is performed in the following way: we randomly generate a correlation matrix C and diagonal matrix Σ^2 with diagonal elements having distribution $\mathrm{U}[0.5, 10]$. Then we compute the covariance matrix $V = \Sigma C \Sigma$, its Cholesky decomposition $V = S^\top S$ and finally set $y_i = Sz_i + \mu$, where $\mu := (0, \dots, d-1)$. We consider $N = 10000$ samples together with $n = 100$ observations. The loss function equals to

$$\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{d} |\hat{x}_{i,j} - \mu_j|, \tag{3.11}$$

where $\hat{x}_{i,j}$ denotes an estimate for sample i and coordinate $j = 1, \dots, d$.

We present the results in Table 3.2 for $X = \mathbb{R}$ and in Table 3.3 for $X = \mathbb{R}^{15}$. The values show the loss function (3.11) for contaminating distribution (first column) with given probability (second column). The bold numbers are the best values among all estimators and the numbers in italic are those within 5% of the best loss function value. For $X = \mathbb{R}$, the results between M-estimators and our estimators are comparable with M-estimators in a slight lead, which is not surprising due to Remark 3.1.12. For $X = \mathbb{R}^{15}$, the performance of both estimators turn around and our estimators perform now better than the examined M-estimators. This is again expected as our estimators try to take into account the covariance structure as explained in Remark 3.1.12. For the second case, our estimators manage to beat the Stahel-Donoho estimators almost in all cases.

Table 3.2. Value of loss function (3.11) for contaminations of $\mathrm{N}(0,1)$ by some other distribution with probability p for $X = \mathbb{R}$.

Distribution	Mean	Med	Trun	Winsor	M1	M2	GM1	GM2	GM3
$p = 0\%$	**0.080**	0.099	0.085	*0.083*	*0.081*	*0.081*	0.116	0.085	*0.083*
$p = 5\%$									
$\mathrm{N}(0,100)$	0.190	0.103	0.089	0.089	**0.084**	*0.088*	0.119	*0.087*	*0.086*
Cauchy	0.657	0.099	*0.086*	*0.084*	**0.082**	*0.083*	0.116	*0.086*	*0.084*
$\mathrm{U}(-10,10)$	0.129	0.103	0.090	*0.089*	**0.084**	*0.088*	0.120	*0.087*	0.086
$p = 10\%$									
$\mathrm{N}(0,100)$	0.259	0.110	0.096	0.097	**0.089**	0.098	0.125	*0.090*	*0.089*
Cauchy	0.642	0.101	*0.088*	*0.087*	**0.084**	*0.086*	0.116	*0.088*	*0.086*
$\mathrm{U}(-10,10)$	0.162	0.108	0.095	0.096	**0.090**	0.096	0.123	*0.090*	*0.090*
$p = 40\%$									
$\mathrm{U}(-20,20)$	0.572	0.163	0.212	0.407	0.182	0.236	*0.150*	*0.149*	**0.147**

Table 3.3. Value of loss function (3.11) for contaminations of $\mathrm{N}(0,I)$ by some other distribution with probability p for $X = \mathbb{R}^{15}$.

Distribution	Mean	Med	M1	SD	GM1	GM2	GM3
$p = 0\%$	**0.151**	*0.154*	*0.151*	0.159	0.213	*0.151*	*0.152*
$p = 5\%$							
$\mathrm{N}(0,100)$	0.360	*0.162*	**0.155**	*0.160*	0.215	*0.155*	*0.156*
Cauchy	2.424	*0.160*	*0.160*	*0.160*	0.214	**0.155**	*0.158*
$\mathrm{U}(-10,10)$	0.241	*0.161*	*0.160*	*0.160*	0.215	**0.155**	*0.157*
$p = 10\%$							
$\mathrm{N}(0,100)$	0.491	0.169	*0.160*	*0.161*	0.217	**0.159**	*0.160*
Cauchy	2.095	*0.167*	0.171	*0.162*	0.216	**0.160**	*0.165*
$\mathrm{U}(-10,10)$	0.308	0.168	0.171	*0.162*	0.217	**0.159**	*0.162*
$p = 40\%$							
$\mathrm{U}(-20,20)$	1.468	0.330	0.292	**0.261**	0.303	0.278	0.280

We depict again the loss function (3.11) for a distribution without any contamination in Table 3.4. Note that all distributions are symmetric with respect to zero. It is visible, that for distributions with heavy tails the more robust estimators perform better. We omitted the M-estimators.

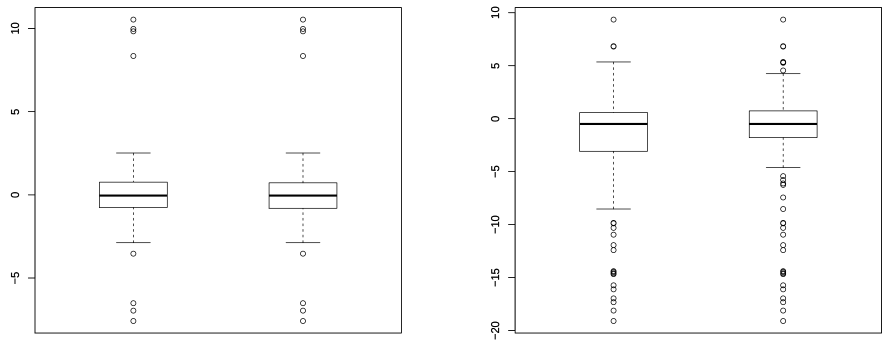
Table 3.4. We depict the value of loss function (3.11) corresponding to various distributions with no contamination in this table. We have $N = 1000$ samples with $n = 100$ observations.

Distribution	Mean	Med	Trun	Winsor	M1	M2
$\mathrm{N}(0,1)$	**0.080**	0.099	0.085	*0.083*	*0.081*	*0.081*
Cauchy	6.150	**0.126**	0.136	0.166	0.140	0.154
t_5	*0.102*	0.104	*0.094*	*0.094*	*0.093*	**0.093**
t_{10}	0.088	0.103	*0.089*	*0.088*	*0.087*	**0.087**
$\mathrm{U}(-10,10)$	**0.461**	0.780	0.610	0.551	0.491	*0.467*
Laplace	0.112	**0.086**	0.093	0.101	0.099	0.097

	GM1	GM2	GM3
$\mathrm{N}(0,1)$	0.116	0.085	*0.083*
Cauchy	*0.128*	0.139	*0.130*
t_5	0.118	*0.095*	*0.093*
t_{10}	0.120	0.092	*0.089*
$\mathrm{U}(-10,10)$	0.983	*0.461*	0.521
Laplace	0.092	*0.099*	*0.095*

3.3.2 Relation to boxplot

Boxplot was proposed for the first time in (Tukey, 1977). It takes the median, then computes the interquartile range (IQR), which is later widened. The observations which are not present in this widening (known as whiskers) are considered as outliers. We compare boxplot with our method, where instead of considering IQR, we take $\boldsymbol{y} \in \mathcal{L}(\hat{T}_n(\boldsymbol{x}), \boldsymbol{x})$. Thus, we do not need to take 25% of the observations with a lower value than the median and 25% of the observations with higher value than the median, but we take 50% of the observations closest to the median. Whiskers are based on $\mathcal{E}_{\boldsymbol{b}}(\{\boldsymbol{y}\}, \boldsymbol{x})$.



(a) $p = 0.1$, $n = 100$, $\mathrm{N}(0, 100)$

(b) $p = 0.4$, $n = 100$, $\mathrm{U}(-20, 10)$

Figure 3.1. Comparison of the classical boxplot (left-hand side of each figure) and our modification (right-hand side of each figure). In both cases we contaminate $\mathrm{N}(0, 1)$ with a probability p by the distribution described under the figures.

We depict this comparison in Figure 3.1. We contaminate the standard normal distribution by some other distribution. In each (sub)figure, the left-hand side is the boxplot and the right-hand side is our modification. Since both approaches differ in the way in which they treat non-symmetry, both graphs in the left figure are identical. However, if we assume non-symmetric distributions (right figure), then our modification is able to detect outliers in a better way than the standard boxplot.

3.3.3 Relation to bagplot

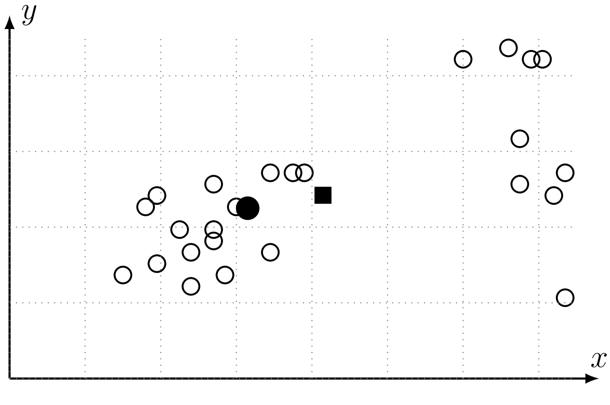
We consider generalization of boxplot into more dimensions in this subsection. This is known as bagplot for two dimensions and was studied for the first time in (Rousseeuw et al., 1999). For generalization to functional data, see, e.g., (Sun and Genton, 2011).

To construct bagplot, a real number called depth is assigned to every observation, see (Zuo and Serfling, 2000). Then the observation with the highest value of depth is called the depth median and the convex hull of approximately 50% of the observations with the highest depth is called the bag (this corresponds to IQR for boxplot). Then the boundary of the bag is enlarged to the polygon called fence (which corresponds to whiskers for boxplot). The area between the fence and the bag is called the loop. The observations not in the fence are considered as outliers.

We present now our modification of the bagplot, illustrated by pictures.

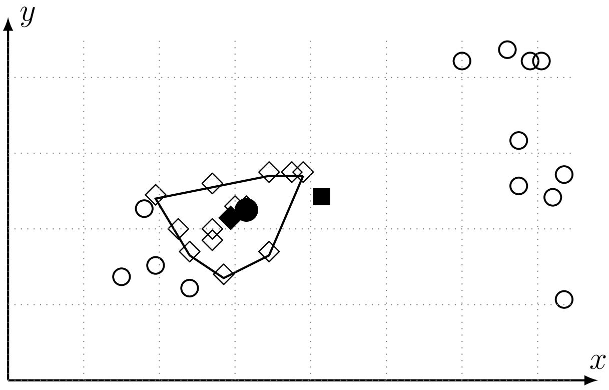
1. Having a sample $\boldsymbol{x}$ from two different distributions, we compute first its geometric median $\hat{T}_n(\boldsymbol{x})$. For simplicity, we assume that it is defined uniquely. Here, the geometric median corresponds to the depth median.

○ Observation, ■ Mean, ●$\hat{T}_n(\boldsymbol{x})$.



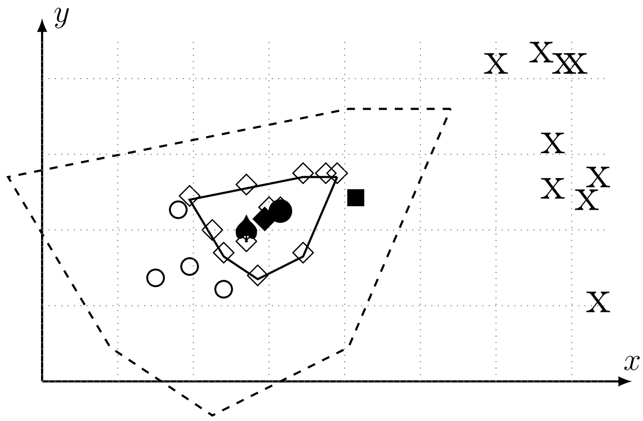
2. Then we choose an arbitrary $\boldsymbol{y} \in X^{\lfloor \frac{n-1}{2} \rfloor}$ from $\mathcal{L}(\hat{T}_n(\boldsymbol{x}), \boldsymbol{x})$, construct a convex hull containing all coordinates of $\boldsymbol{y}$ and call this set the bag. We denote, as in the previous section, a mean of observations from $\boldsymbol{y}$ by GM1.

—Convex hull of points in $\boldsymbol{y}$, ◆ GM1.



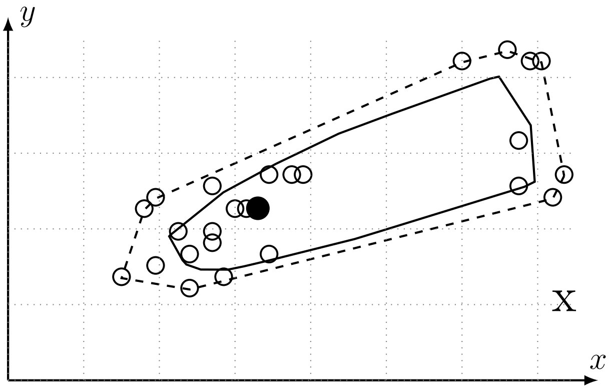
3. In the last step, we enlarge the bag by the following procedure: fix a constant $a = 3$, denote by V the set of all coordinates of $\boldsymbol{y}$ and define the fence as
$$\operatorname*{conv}_{v \in V}\{\hat{T}_n(\boldsymbol{x}) + a(v - \hat{T}_n(\boldsymbol{x}))\},$$
where conv stands for the convex hull. Then we denote the mean of all observation in the fence by GM2.

-- Boundary of the set E, ♠ GM2, X Outlier.



4. The last picture depicts the original solution described in (Rousseeuw et al., 1999). We do not compute the values, instead of it we utilize the procedure **bagplot()** from the package aplpack of the software R. We see in the picture that the loop and also the bag covers the observations from both subsets of observations given by different distributions.

● Depth median, — Boundary of the bag, -- Fence.



We will comment briefly on the last step of the algorithm. Constant a was chosen in the same way which was recommended in (Rousseeuw et al., 1999). Even though the construction of the fence is similar to $\mathcal{E}_{\boldsymbol{b}}(\mathcal{L}(\hat{T}_n(\boldsymbol{x}),\boldsymbol{x}),\boldsymbol{x})$, the inflating is performed in a slightly different way.

Moreover, since the construction of GM2 is based on objects which have the limiting breakdown point of $\frac{1}{2}$, estimate GM2 will have the same property. On the other hand, since bagplot is a generalization of boxplot, which has the breakdown point of $\frac{1}{4}$, we cannot expect the bagplot to have higher breakdown point than this. Thus, our version of bagplot deals better with outliers than the original method.

3.3.4 Outlying shortfall

Risk measures are essential for banks as a tool for measuring the risk embedded in its portfolios. The most pronounced concept is known as coherent risk measures. It was proposed in (Artzner et al., 1999). In order to introduce the concept, we have to give a formal definition of risk measure. Let us have some time horizon T. Let $L(\Omega,\mathcal{A},\mathrm{P})$ denote the set of all almost surely finite random variables on $(\Omega,\mathcal{A})$. Further, we employ $\mathcal{M}\subset L(\Omega,\mathcal{A},\mathrm{P})$ the set of random variables interpreted as a potential loss of an investigated portfolio over the horizon T. We assume $\mathcal{M}$ to be convex cone, i.e., that $L_1,L_2\in\mathcal{M}$ implies that $L_1+L_2\in\mathcal{M}$ and $\lambda L_1\mathcal{M}$ for every $\lambda>0$. Risk measure is the function $\rho:\mathcal{M}\to\mathbb{R}$. The coherent risk measure has to fulfill the following properties.

1. (translation invariance) For all $L\in\mathcal{M}$ and every $l\in\mathbb{R}$ holds $\rho(L+l)=\rho(L)+l$.

2. (subadditivity) For all $L_1,L_2\in\mathcal{M}$ we have $\rho(L_1+L_2)\leq\rho(L_1)+\rho(L_2)$.

3. (positive homogeneity) For all $L\in\mathcal{M}$ and $\lambda>0$ we have $\rho(\lambda L)=\lambda\rho(L)$.

4. (monotonicity) For L_1,L_2 such that $L_1\leq L_2$ almost surely it holds $\rho(L_1)\leq\rho(L_2)$.

One of the most famous risk measures are value at risk (VaR) and excpected shortfall (ES) also known as conditional value at risk.

Definition 3.3.1. *Let $\alpha\in(0,1)$ and $L\in\mathcal{M}$*

$$\mathrm{VaR}_\alpha(L)=\inf\left\{l\in\mathbb{R}\,|\,\mathrm{P}(L>l)\geq 1-\alpha\right\}.$$

We define expected shortfall with the help of VaR.

Definition 3.3.2. *Let $\alpha \in (0,1)$ and $L \in \mathcal{M}$*

$$\mathrm{ES}_\alpha(L) = \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{VaR}_u(L)du.$$

Expected shortfall is coherent measure as is demonstrated in (Embrechts et al., 2005). However, value at risk is not, because it does not fulfill the subadditivity axiom.

We can construct a similar statistic to expected shortfall.

Definition 3.3.3. *Let $\alpha \in (0,1)$, $L \in \mathcal{M}$, F_L be a distribution function of L and*

$$a(L) = \inf_{b>0}\left(\mathrm{P}(\|L - \hat{T}(L)\| \leq b) \geq \frac{1}{2}\right).$$

Compare the definition of $a(L)$ with (3.10). *We define*

$$\mathrm{OS}_\alpha(L) = \frac{1}{1-\beta(L)}\int_\beta^1 \mathrm{VaR}_u(L)du, \tag{3.12}$$

where $\beta(L) = 1 - F_L\left(q_{\mathrm{N}(0,1)}(1-\alpha)\frac{a(L)}{q_{\mathrm{N}(0,1)}(0.75)-q_{\mathrm{N}(0,1)}(0.25)} + \hat{T}(L)\right)$ and $q_{\mathrm{N}(\mu,\sigma)}(\alpha)$ is an α quantile of $\mathrm{N}(\mu,\sigma)$.

Equation (3.12) can be rewritten in the following way

$$\mathrm{OS}_\alpha(L) = \frac{1}{1-\beta(L)}\int_{b(L)}^\infty u dF_L(u),$$

where $b(L) = q_{\mathrm{N}(0,1)}(1-\alpha)\frac{a(L)}{q_{\mathrm{N}(0,1)}(0.75)-q_{\mathrm{N}(0,1)}(0.25)} + \hat{T}(L)$.

This can be interpreted as an average of observations which are greater than $\alpha-$quantile of the normal distribution $\mathrm{N}(\hat{\mu}, \hat{\sigma}^2)$, where $\hat{\mu}$ and$\hat{\sigma}$ are robust estimators of expected value and standard deviation of L.

Example 3.3.4. *We want to deal with the coherence of outlying shortfall in this example. We show that the properties of subadditivity and monotonicity are not fulfilled.*

Consider two correlated variables L_1 and L_2.

$$[L_1, L_2] = \begin{cases} [q_{\mathrm{N}(0,1)}(0.25), q_{\mathrm{N}(0,1)}(0.25)] & \textit{with probability } \frac{2}{5}, \\ [q_{\mathrm{N}(0,1)}(0.75), q_{\mathrm{N}(0,1)}(0.75)] & \textit{with probability } \frac{2}{5}, \\ [q_{\mathrm{N}(0,1)}(1-\alpha)-1, q_{\mathrm{N}(0,1)}(1-\alpha)+2] & \textit{with probability } \frac{1}{5}. \end{cases}$$

It is clear that for $a(L_1) = a(L_2) = q_{\mathrm{N}(0,1)}(0.75) - q_{\mathrm{N}(0,1)}(0.25)$. It also follows $\mathrm{OS}_\alpha(L_1) = 0$ from the definition but $\mathrm{OS}_\alpha(L_2) = \frac{1}{5}q_{\mathrm{N}(0,1)}(1-\alpha)+2$. We also get $a(L_1+L_2) = 2(q_{\mathrm{N}(0,1)}(0.75) - q_{\mathrm{N}(0,1)}(0.25))$. I.e., we have to consider all values greater than $2q_{\mathrm{N}(0,1)}(1-\alpha)$ with appropriate probabilities. This gives $\mathrm{OS}_\alpha(L_1 + L_2) = 2q_{\mathrm{N}(0,1)}(1-\alpha)+1 > q_{\mathrm{N}(0,1)}(1-\alpha)+2 = \mathrm{OS}_\alpha(L_1)+\mathrm{OS}_\alpha(L_2)$. Therefore, the axiom of subadditivity does not hold.

We focus now our attention on monotonicity. Once more we consider two correlated random variables

$$[L_1, L_2] = \begin{cases} [q_{\mathrm{N}(0,1)}(0.25) + 2, q_{\mathrm{N}(0,1)}(0.25)] & \textit{with probability } \frac{2}{5}, \\ [q_{\mathrm{N}(0,1)}(0.75) + 2, q_{\mathrm{N}(0,1)}(0.75)] & \textit{with probability } \frac{2}{5}, \\ [q_{\mathrm{N}(0,1)}(1-\alpha) + 1, q_{\mathrm{N}(0,1)}(1-\alpha) + 1] & \textit{with probability } \frac{1}{5}. \end{cases}$$

From this follows that $L_1 \geq L_2$ almost surely; however, $OS_\alpha(L_1) = 0$ and $OS_\alpha(L_2) = q_{\mathrm{N}(0,1)}(1-\alpha)+1$. Therefore, the axiom of monotonicity is not fulfilled.

Let us deal with positive homogeneity. $F_{\lambda L}(x) = \mathrm{P}(\lambda L \leq x) = F_L(\frac{x}{\lambda})$. We have

$$a(\lambda L) = \inf_{b>0} \left(\mathrm{P}(\|\lambda L - \hat{T}(\lambda L)\| \leq b) \geq \frac{1}{2} \right) = \inf_{b>0} \left(\mathrm{P}(\|L - \hat{T}(L)\| \leq \frac{b}{\lambda}) \geq \frac{1}{2} \right),$$

where we utilize the scale equivariance of the median. It yields $a(\lambda L) = \lambda a(L)$ which gives $b(\lambda L) = \lambda b(L)$. It holds $F_{\lambda L}(x) = \mathrm{P}(\lambda L \leq x) = F_L(\frac{x}{\lambda})$. This gives $\beta(\lambda L) = 1 - F_L\left(\frac{b(\lambda L)}{\lambda}\right) = 1 - F_L(b(L)) = \beta(L)$. At the end we compute

$$\begin{aligned} \mathrm{OS}_\alpha(\lambda L) &= \frac{1}{1-\beta(\lambda L)} \int_{b(\lambda L)}^{\infty} u dF_{\lambda L}(u) = \frac{1}{1-\beta(L)} \int_{\lambda b(L)}^{\infty} u dF_L\left(\frac{u}{\lambda}\right) \\ &= \frac{\lambda}{1-\beta(L)} \int_{b(L)}^{\infty} u dF_L(u) = \lambda\, \mathrm{OS}_\alpha(L), \end{aligned}$$

where we employ substitution.

The axiom of translation invariance can be proved in a similar way. △

We have shown in Example 3.3.4 that the theoretical properties of the new risk measure are not satisfactory. Nevertheless, it can still serve as a measure of unexpected losses. This measure is also slightly more complicated and therefore its interpretation can be more difficult.

The normal quantiles can be replaced by some other distribution in the case that we suppose that the random variable L has this distribution potentially contaminated by some other distribution.

We employ EUR/USD hourly rates from 1. 8. 2012 to 3. 11. 2012 as a small illustration of this approach.

Let r_t be a logarithmic return in time t. We denote the realization of the loss L in time t as $L_t = -r_t$. We suppose that r_t are independent and have a normal distribution contaminated by some other distribution. We are interested in the value of potential loss due to contamination.

Results:

- 95 % VaR is 0.00140406.
- 95 % expected shortfall is 0.00219346.
- Outlying shortfall (with a constant of widening with $\alpha = 0.01$) is 0.002376221.
- 99 % VaR is 0.002517572.
- 99 % expected shortfall is 0.003608807.

We can also determine how often we suffer a loss because of outlying observations. In our case it is $\beta(L) = 3.91\%$.

3.3.5 Multidimensional value at risk

We have introduced outlying shortfall in the previous section. We look now at one possible way how to generalise VaR. Modern approaches to multidimensional value at risk can be found e.g., in (Prékopa, 2012). We just want to hint at a different way, based on an idea from (Koenker and Bassett, 1978), to compute multidimensional quantiles. We combine the approaches of Chapters 2 and 3 in this way.

As the first step, we describe the data we want to deal with. They are gained by simulation study, of which details are described later. Let us assume that we have two series of daily returns $r_{i,t}$ of some stocks S_i, where $i = 1, 2$ and $t = 1, \dots, T$. We suppose that the returns are independent for each time (they are not necessarily independent in one day between each other). We are interested in daily (for simplicity) value at risk ($\mathrm{VaR}_\alpha(L_i)$) at a confidence level $\alpha = 5\%$. The random variable L_i expresses the loss from the stock S_i. Therefore, $L_{i,t} = -r_{i,t}$. Let us assume that we have three years of history which means $T = 800$ observations.

We consider a linear dependence between the two returns. The dependence is measured by Pearson covariance matrix $\hat{\Sigma}$. Cholesky decomposition $\hat{\Sigma} = CC^\top$ is applied on the matrix $\hat{\Sigma}$, in the case of $\hat{\Sigma}$ regular. Let $\boldsymbol{r}_t = (r_{1,t}, r_{2,t})^\top$. We compute $\boldsymbol{y}_t^\top = \boldsymbol{r}_t^\top C^{-1}$ and put $Y = (\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_T)^\top$ as a $2 \times T$ matrix. We apply the following formula

$$\hat{\boldsymbol{r}}_\alpha^M = C \operatorname*{argmin}_{\boldsymbol{y} \in \mathbb{R}^d} \sum_{t=1}^{T} \rho_\alpha^M(\boldsymbol{y}_t - \boldsymbol{y}), \tag{3.13}$$

where

$$\rho_\alpha^M(\boldsymbol{x}) = \|(|y_1|\{\alpha \mathsf{I}_{[y_1 \geq 0]} + (1-\alpha)\mathsf{I}_{[y_1 < 0]}\}, \dots, |y_d|\{\alpha \mathsf{I}_{[y_d \geq 0]} + (1-\alpha)\mathsf{I}_{[y_d < 0]}\})\|.$$

Regarding our example $d = 2$. We put $\mathrm{VaR}_\alpha(L_1, L_2) = -\hat{\boldsymbol{r}}_\alpha^M$.

We multiply the vector resulting from minimization by C in Equation (3.13) to get the same dependence structure as in the data.

If we omit the usage of C and compute

$$\hat{\boldsymbol{r}}_\alpha^M = \operatorname*{argmin}_{\boldsymbol{r} \in \mathbb{R}^d} \sum_{t=1}^{T} \rho_\alpha^M(\boldsymbol{r}_t - \boldsymbol{r}), \tag{3.14}$$

instead of (3.13) then the solution would not take into consideration the dependence structure and so, in the case of positive correlation, the $\mathrm{VaR}_\alpha(L_1, L_2)$ would be too low and, in the case of negative correlation, too high.

We describe now the details of the simulation. We simulate $r_{1,t}$ from standard normal distribution $\mathrm{N}(0, 1)$ contaminated by t_4 distribution with a probability of 5 %. The series $r_{2,t}$ has a standard normal distribution $\mathrm{N}(0, 1)$ contaminated by t_6 distribution with a probability of 10 %. Further, we consider the first series with twice higher weight i.e., we multiply the first series of the returns by two. This is motivated by the different representations of stocks in a usual portfolio.

Since we have only two series, we consider the correlation coefficient between them as ρ.

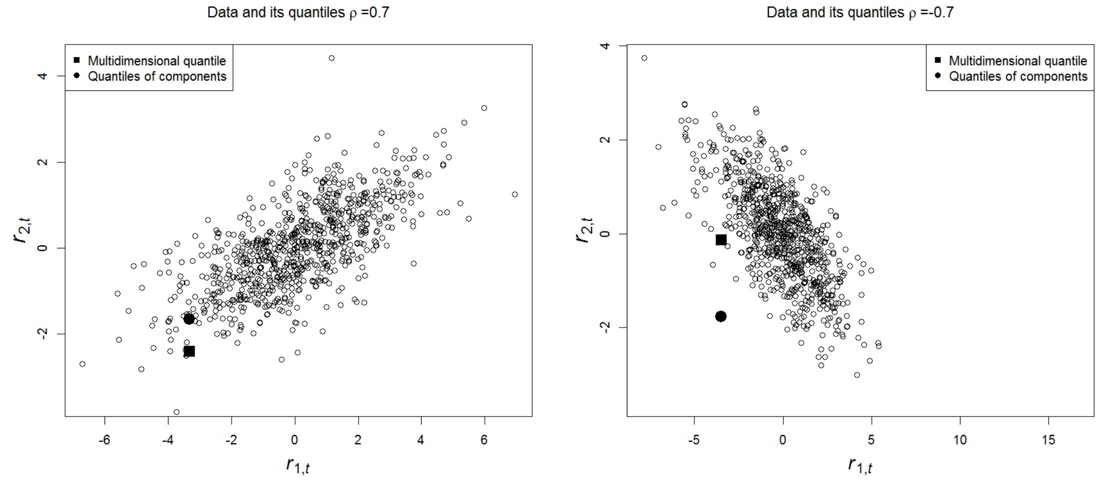


Figure 3.2. The depicted figures are time series simulated with different correlation ρ. We have 800 observations. We estimate 5 % quantile by our method (see (3.13)) and componentwise (see (3.14)).

The first part of Picture 3.2 depicts the situation when correlation is positive. We get for the second component a very low value. I.e., when we have lost in the first component we should await a loss in the second as well. The next part of the picture demonstrates an opposite situation. We see that in the case of negative correlation the result is very low for the first component and for the second is almost positive. I.e., according to our method: when a bad scenario takes place and we lose in the first component, the loss in the second component would not be so bad. We can also compare the situations when we employ the Equation (3.13) (black box) or (3.14) (black circle).

3.4 Choice of b

In this short section we derive an estimate for b for algorithms GM2 and GM3 described in Table 3.1. Note that due to the construction of the algorithm, it is sufficient to define $b_i := b(x_i, \boldsymbol{x})$ for all observations in $\mathcal{L}(\hat{T}_n(\boldsymbol{x}), \boldsymbol{x})$. Since we want to keep GM2 as simple as possible, we consider constant b and relax this assumption for GM3. Moreover, we derive a different value for one- and more-dimensional cases. We start with a technical lemma.

Lemma 3.4.1. *Let Y be a random variable with a finite second moment, distribution function G, mean μ, standard deviation σ and let $q_{\mu,\sigma}$ denote its quantile function. Then for a fixed $\alpha \in [0,1]$, the following ratio does not depend on the values of μ and σ*

$$K_{G,\alpha} = \frac{q_{\mu,\sigma}(1-\alpha/2) - q_{\mu,\sigma}(\alpha/2)}{q_{\mu,\sigma}(0.75) - q_{\mu,\sigma}(0.25)}.$$

Proof. This follows from the fact $\sigma q_{0,1}(\alpha) + \mu = q_{\mu,\sigma}(\alpha)$. □

For the case of one dimension, consider a random sample $\boldsymbol{x}$ from $\mathrm{N}(\mu, \sigma^2)$ and denote its median by $\hat{T}_n(\boldsymbol{x})$. Assume for simplicity that $\mathcal{L}(\hat{T}_n(\boldsymbol{x}), \boldsymbol{x}) \subset \mathbb{R}^{\left\lfloor \frac{n-1}{2} \right\rfloor}$ is a singleton. Then we may estimate $q_{\mu,\sigma}(0.75) - q_{\mu,\sigma}(0.25)$ by

$$\max \mathcal{L}(\hat{T}_n(\boldsymbol{x}), \boldsymbol{x}) - \min \mathcal{L}(\hat{T}_n(\boldsymbol{x}), \boldsymbol{x}),$$

Then we define b as

$$b := K_{\mathrm{N}(0,1),\alpha} \frac{\max \mathcal{L}(\hat{T}_n(\boldsymbol{x}), \boldsymbol{x}) - \min \mathcal{L}(\hat{T}_n(\boldsymbol{x}), \boldsymbol{x})}{2} \asymp \frac{q_{\mu,\sigma}(1-\alpha/2) - q_{\mu,\sigma}(\alpha/2)}{2}.$$

This value thus estimates the distance between chosen quantiles divided by two. If we take all observations of which the distance from median is b, then we get, under the assumption of uncontaminated normality, approximately $1-\alpha$ of our observations.

For the multidimensional case $\mathbb{R}^d$, we assume that $\boldsymbol{x}$ follows the $\mathrm{N}(\mu, \sigma^2 I)$ distribution. We use $\max_{y \in \mathcal{L}(\hat{T}_n(\boldsymbol{x}),\boldsymbol{x})} \|y - \hat{T}_n(\boldsymbol{x})\|$ as an approximation of $\sigma \chi^2_{0.5,d}$, where $\chi^2_{0.5,d}$ is the quantile function of χ^2 distribution with d degrees of freedom evaluated at probability 0.5. To include approximately fraction $\alpha \in (0,1)$ of all observations, we set

$$b := \max_{y \in \mathcal{L}(\hat{T}_n(\boldsymbol{x}),\boldsymbol{x})} \|y - \hat{T}_n(\boldsymbol{x})\| \frac{\chi^2_{\alpha,d}}{\chi^2_{0.5,d}}.$$

For GM3, we consider directly $\mathbb{R}^d$. Assume again that our sample $\boldsymbol{x}$ has distribution $\mathrm{N}(\mu, \sigma^2 I)$, then for x_i from the boundary of $\mathcal{L}(\hat{T}_n(\boldsymbol{x}), \boldsymbol{x})$ we have

$$\frac{\|x_i - \hat{T}_n(\boldsymbol{x})\|}{\sigma} \sim \sqrt{\chi^2_{0.5,d}},$$

Now, fixing given probability level $\alpha \in (0,1)$ and weight $w \in (0,1)$, we want to have all x_e with

$$\frac{\|x_e - \hat{T}_n(\boldsymbol{x})\|}{\sigma} \sim \sqrt{\chi^2_{\alpha,d}}$$

to have weight (3.7) at least w. But plugging this in the definition of weight results in

$$\begin{aligned} b_i &\geq \frac{\|x_i - x_e\|}{1-w} \geq \frac{\|x_e - \hat{T}_n(\boldsymbol{x})\|}{1-w} - \frac{\|x_i - \hat{T}_n(\boldsymbol{x})\|}{1-w} \sim \frac{\sigma}{1-w}\left(\sqrt{\chi^2_{\alpha,d}} - \sqrt{\chi^2_{0.5,d}}\right) \\ &= \frac{\sigma}{1-w} \frac{\|x_i - \hat{T}_n(\boldsymbol{x})\|}{\|x_i - \hat{T}_n(\boldsymbol{x})\|}\left(\sqrt{\chi^2_{\alpha,d}} - \sqrt{\chi^2_{0.5,d}}\right). \end{aligned}$$

Approximating again the distribution and taking minimum value of b_i, we set

$$b_i := \frac{\sqrt{\chi^2_{\alpha,d}} - \sqrt{\chi^2_{0.5,d}}}{(1-w)\sqrt{\chi^2_{0.5,d}}} \|x_i - \hat{T}_n(\boldsymbol{x})\|.$$

In the numerical experiments, we have chosen $\alpha = 0.99$.

4. Conclusions: reached results and possible future work

We want to summarize our results and compare suggested methods in this section. We will also mention potential improvements and suggest the direction of future research. It consist of deeper investigation of theoretical properties and also improvements in performance of our algorithms.

We offered new robust algorithms based on the approach connected to median in this dissertation thesis. We studied these methods with respect to breakdown point, consistency and computationally complexity.

Chapter 1 was split into two parts. In the first part we introduced the background of the studied problems simultaneously with relevant literature.

In the second part we introduced the notation, definitions and basic concepts utilised throughout our work. We discussed possible adjustments for definition of breakdown point especially for recursive adaptive algorithms. We also considered the computation of probability that the smoothed value of the series is influenced by outliers. These new concepts were accompanied by examples.

The recursive adaptive algorithms were dealt in Chapter 2. We proposed robust alternatives to exponential smoothing by the description of two different concepts.

- The first approach exploits the properties of absolute norm and exponential weighting. It is basically very similar to exponential smoothing and it was already studied. Nevertheless, we generalised it for quantiles i.e., we extended the algorithm for a broader class of objective functions. We also attempted to generalize the method to the more dimensional case, but our suggested method is too complicated.

- The other approach exploits sign test. With its help we test whether there is a change point in a series. The method was generalised to the more dimensional case. Unfortunately, it does not have satisfactory convergence properties. However, possible improvements were proposed.

These concepts cannot be easily generalized for ARMA processes.

The chapter is finished by simulation study and the example on real data. We have studied the methods from the perspective of breakdown point, computational complexity and consistence.

Chapter 3 deals with new estimators of location based on geometric median, which can be also interpreted as generalisation of weighted trimmed mean. In the main part of the chapter general vector space is considered and the estimators are regarded as sets. We also briefly mention a relation to the concept of depth. The method is compared by simulation study and is utilized in construction of boxplot and bagplot. We also proposed new measures of risk. However, they can serve only as an illustration of our approach, because the substantial improvements concerning their properties are necessary.

The further work can contain better determination of critical values in the case of the sign test algorithms. They can also be generalized to look for quantiles. Another possible generalization includes seasonality. We could also employ different estimator of location than median or employ another test. As we have mentioned the field of robust sequential change detection tests is quite vast. Therefore, nice work can be done if we compare them with respect to our assumptions and utilisation. Also a research concerning the speed of convergence would be feasible.

Potential for future research is also offered in time series analysis. In this area were published articles e.g., (Zielinski, 1999) and (Luger, 2006). However, they consider only the case AR(1).

We have mentioned that this dissertation is focused on simple algorithms based on median with respect to its robust properties. This theme is incorporated throughout our work. However, there are many different results without direct connection to each other. Especially the connection between Chapters 2 and 3 is not obvious. We have mentioned in the beginning of Chapter 3 that the results of the chapter can be utilized in Chapter 2. Nevertheless, it would demand an addition of another chapter connecting the previous ones and this is left for future research.

Potential improvements are also possible concerning the weighting function of Algorithm 3.1.1. Its convergence properties should be studied. However, the consistence of the other proposed methods is studied, but also here a deeper analysis can still be applied.

Table 4.1. In this table we summarise our results. It serves only for rough idea.

Method	Equivariance	Convergence	Robustness	Complexity
C-algorithm	Shift, scale ($\alpha = 0.5$)	No	$\frac{j}{W}$ $1 \leq j \leq W$	$\mathcal{O}(n)$
Sign test alg.: Constant trend	Shift, scale	The more obs. the better	$\frac{1}{2}$ of $m(t, \boldsymbol{y})$	$\mathcal{O}(n^2)$
Sign test alg.: Linear trend	Shift, scale	No	$\frac{1}{2}$ of $m(t, \boldsymbol{y})$	$\mathcal{O}(n^2)$
Sign test alg.: Linear trend, modification	Shift, scale	Fisher consistent	$\frac{1}{2}$ of $m(t, \boldsymbol{y})$	$\mathcal{O}(n^2 \log n)$
Trimmed mean	Shift, scale	Weak	$\frac{1}{2}$	$\mathcal{O}(n \log n)$

Table 4.1 serves as a rough summary of our results. We employ there the notation, where n is the number of observations, $W \in \mathbb{N}$ denotes the number of observations employed for initial estimate and $m(t, \boldsymbol{y})$ the number of observations on which base is computed a smooth value for observation y_t.

Bibliography

Artzner, P., Delbaen, F., Eber, J. M. and Heath, D. (1999): 'Coherent measures of risk', *Mathematical finance* **9**(3), 203–228.

Brönnimann, H. and Chazelle, B. (1998): 'Optimal slope selection via cuttings', *Computational Geometry Theory and Applications* **10**(1), 23–29.

Brown, R. G. (1962): *Smoothing, Forecasting and Prediction of Descrete Time Series*, Prentice-Hall, Englewood Cliffs, NJ.

Cadre, B. (2001): 'Convergent estimators for the L_1-median of a Banach valued randomvariable', *Statistics* **35**(4), 509–521.

Cardot, H., Cénac, P. and Zitt, P. A. (2013): 'Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm', *Bernoulli* **19**(1), 18–43.

Cipra, T. (1992): 'Robust exponential smoothing', *Journal of Forecasting* **11**(1), 57–69.

Cipra, T. and Romera, R. (1992): 'Recursive time series methods in L_1 - norm', *In: L_1-Statistical Analysis and Related Methods (Y. Dodge ed.)* pp. 233–243.

Cormen, T. H. (2009): *Introduction to Algorithms*, MIT press.

Donoho, D. L. (1982): Breakdown Properties of Multivariate Location Estimators, *in* 'Ph.D. qualifying paper', Harvard University.

Donoho, D. L. and Huber, P. J. (1983): The Notion of Breakdown Point, *in* P. J. Bickel, K. A. Doksum and J. L. Hodges, eds, 'A Festschrift for Erich Lehman', Wadsworth, pp. 157–184.

Embrechts, P., Frey, R. and McNeil, A. (2005): 'Quantitative risk management', *Princeton Series in Finance, Princeton* **10**.

Gelper, S., Fried, R. and Croux, C. (2010): 'Robust forecasting with exponential and Holt–Winters smoothing', *Journal of Forecasting* **29**(3), 285–300.

Genton, M. G. and Lucas, A. (2003): 'Comprehensive definitions of breakdown points for independent and dependent observations', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**(1), 81–94.

Gervini, D. (2008): 'Robust functional estimation using the median and spherical principal components', *Biometrika* **95**(3), 587–600.

Haldane, J. B. S. (1948): 'Note on the median of a multivariate distribution', *Biometrika* **35**(3-4), 414–417.

Hampel, F. R. (1971): 'A general qualitative definition of robustness', *The Annals of Mathematical Statistics* **42**(6), 1887–1896.

Hanxiang, P., Shaoli, W. and Xueqin, W. (2008): 'Consistency and asymptotic distribution of the theil–sen estimator', *Journal of Statistical Planning and Inference* **138**(6), 1836–1850.

Hanzák, T. and Cipra, T. (2011): 'Exponential smoothing for time series with outliers', *Kybernetika* **47**(2), 165–178.

Hayford, J. F. (1902): 'What is the center of an area, or the center of a population?', *Publications of the American Statistical Association* **8**(58), 47–58.

Holt, C. C. (2004): 'Forecasting seasonals and trends by exponentially weighted moving averages', *International Journal of Forecasting* **20**, 5–10.

Hossjer, O., Rousseeuw, P. J. and Croux, C. (1994): 'Asymptotics of the repeated median slope estimator', *The Annals of Statistics* pp. 1478–1501.

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

Huber, P. J. (1981): *Robust Statistics*, John Wiley & Sons.

Huber, P. J. and Ronchetti, E. M. (2009): *Robust Statistics*, John Wiley & Sons, Inc., Hoboken, NJ, USA.

Hušková, M. (2014): Robust change point analysis, *in* C. Becker, R. Fried and S. Kuhnt, eds, 'Robustness and Complex Data Structures: Festschrift in Honour of Ursula Gather', Springer Science & Business Media, chapter 11, pp. 171–190.

Hušková, M. and Sen, P. K. (1989): *Nonparametric Tests for Shift and Change in Regression at an Unknown Time Point*, Springer-Verlag.

Hyndman, R. J., Koehler, A. B., Ord, J. K. and Snyder, R. D. (2008): *Forecasting with Exponential Smoothing*, Springer, Berlin Heidelberg.

Jurečková, J. and Picek, J. (2005): *Robust Statistical Methods with R*, Chapman and Hall/CRC.

Kalina, J. (2009): 'Least weighted squares in econometric applications', *Journal of Applied Mathematics, Statistics and Informatic* **5**(2), 115–125.

Kemperman, J. H. B. (1987): Note on the Median of a Multivariate Distribution, *in* 'Statistical Data Analysis Based on the L_1-norm and Related Methods', Neuchâtel, North-Holland, Amsterdam, pp. 217–230.

Koenker, R. and Bassett, G. (1978): 'Regression quantiles', *Econometrica* **46**(1), 33–50.

Koubková, A. (2004): Forecasting Seasonals and Trends by Exponentially Weighted Moving Averages, *in* J. Antoch, ed., 'COMPSTAT 2004 Proceedings', Springer Verlag, pp. 1345–1352.

Koubková, A. (2006): Sequential Change-Point Analysis, *in* 'Ph.D. Thesis', Charles University in Prague.

Lehmann, E. L. and Casella, G. (2006): *Theory of Point Estimation*, Springer Science & Business Media.

Liu, R. Y. (1990): 'On a notion of data depth based on random simplices', *Annals of statistics* **18**, 405–414.

Lopuhaä, H. P. (1991): 'Multivariate τ-estimators for location and scatter', *Canadian Journal of Statistics* **19**, 307–321.

Luger, R. (2006): 'Median-unbiased estimation and exact inference methods for first-order autoregressive models with conditional heteroscedasticity of unknown form', *Journal of Time Series Analysis* **27**(1), 119–128.

Maronna, R. (1976): 'Robust M-estimators of multivariate location and scatter', *The Annals of Statistics* **4**, 51–67.

Maronna, R. A., Martin, D. R. and Yohai, V. J. (2006): *Robust Statistics: Theory and Methods*, John Wiley & Sons Ltd, Chichester, West Sussex.

Maronna, R. A. and Zamarb, R. H. (2002): 'Robust estimates of location and dispersion for high-dimensional datasets', *Technometrics* **44**(4), 307–317.

Mašíček, L. (2004): 'Optimality of least weighted squares estimator', *Kybernetika* **40**(6), 715–734.

Mizera, I. and Wellner, J. A. (1998): 'Necessary and sufficient conditions for weak consistency of the median of independent but not identically distributed random variables', *The Annals of Statistics* **26**(2), 672–691.

Moore, G. H. and Wallis, W. A. (1943): 'Time series significance tests based on signs of differences', *Journal of the American Statistical Association* **38**(222), 153–164.

Page, E. S. (1954): 'Continuous inspection schemes', *Biometrika* **41**(1/2), 100–115.

Papageorgiou, M., Kotsialos, A. and Poulimenos, A. (2005): 'Long-term sales forecasting using Holt-Winters and neural network methods', *Journal of forecasting* **24**(5), 353–368.

Pollard, D. (2012): *Convergence of Stochastics Processes*, Springer Science & Business Media.

Polunchenko, A. S. and Tartakovsky, A. G. (2012): 'State-of-the-art in sequential change-point detection', *Methodology and computing in applied probability* **14**(3), 649–684.

Portnoy, S. and Koenker, R. (1997): 'The gaussian hare and the laplacian tortoise: computability of squared-error versus absolute-error estimators', *Statistical Science* **12**(4), 279–300.

Prékopa, A. (2012): 'Multivariate value at risk and related topics', *Annals of Operations Research* **193**(1), 49–69.

Romera, R. and Cipra, T. (1995): 'On practical implementation of robust Kalman filtering', *Communications in Statistics - Simulation and Computation* **24**(2), 461–488.

Rousseeuw, P. J. (1985): Multivariate Estimation With High Breakdown Point, *in* 'Mathematical Statistics and Aplications', Elsevier, Amsterdam, pp. 101–121.

Rousseeuw, P. J., Ruts, I. and Tukey, J. W. (1999): 'The bagplot: a bivariate boxplot', *The American Statistician* **53**(4), 382–387.

Scates, D. E. (1931): *Locating the Median of the Population in the United States*, Instituto Poligrafico Dello Stato.

Seber, G. A. F. and Lee, A. J. (2003): *Linear Regression Analysis*, JOHN WILEY & SONS, New Jersey.

Sen, P. K. (1968): 'Estimates of the regression coefficient based on Kendall's tau', *Journal of the American Statistical Association* **63**, 1379–1389.

Sen, P. K. (1970): 'A note on order statistics for heterogeneous distributions', *The Annals of Mathematical Statistics* **41**(6), 2137–2139.

Serfling, R. (1987): 'Asymptotic behavior of S-estimators of multivariate location parameters and dispersion matrices', *The Annals of Statistics* **15**, 1269–1292.

Siegel, A. F. (1982): 'Robust regression using repeated medians', *Biometrika* **69**(1), 242–244.

Sipser, M. (2012): *Introduction to the Theory of Computation*, Cengage Learning.

Small, C. G. (1990): 'A survey of multidimensional medians', *International Statistical Review/Revue Internationale de Statistique* pp. 263–277.

Sprent, P. and Smeeton, N. C. (2001): *Applied Nonparametric Statistical Methods*, CRC Press.

Stahel, W. A. (1981): Breakdown of Covariance Estimators, *in* 'Research Report 31', Fachgruppe für Statistik, ETH Zürich.

Stigler, S. M. (1973): 'Studies in the history of probability and statistics. XXXII: Laplace, Fisher and the discovery of the concept of sufficiency', *Biometrika* **60**(3), 439–445.

Sun, Y. and Genton, M. G. (2011): 'Functional boxplots', *Journal of Computational and Graphical Statistics* **20**(2), 316–334.

Theil, H. (1950): 'A rank-invariant method of linear and polynomial regression analysis', *Advanced Studies in Theoretical and Applied Econometrics* **23**, 345–381.

Tukey, J. W. (1977): *Exploratory Data Analysis*, Addison-Wesley Publishing Co., Massachusetts.

v. Mises, R. (1947): ‘On the asymptotic distribution of differentiable statistical functions’, *The annals of mathematical statistics* **18**(3), 309–348.

Víšek, J. A. (2000): ‘On the diversity of estimates’, *Computational Statistics and Data Analysis* **34**(1), 67–89.

Víšek, J. A. (2011): ‘Consistency of the least weighted squares’, *Kybernetika* **47**(2), 179–206.

Weber, A. (1909): *Über den Standort der Industrien*, Vol. 2, Pipol Klassik.

Wilcoxon, F. (1945): ‘Individual comparisons by ranking methods’, *Biometrics Bulletin* **1**(6), 80–83.

Wolfowitz, J. (1944): ‘Asymptotic distribution of runs up and down’, *Annals of Mathematical Statistics* **15**(2), 163–172.

Yang, L., Pai, S. and Wang, Y. R. (2010): A novel cusum median control chart, *in* ‘Proceedings of International Multiconference of Engineers and Computer Scientists’, Citeseer.

Zielinski, R. (1999): ‘A median-unbiased estimator of the ar (1) coefficient’, *Journal of Time Series Analysis* **20**(4), 477–481.

Zuo, Y. and Serfling, R. (2000): ‘General notions of statistical depth function’, *The Annals of Statistics* **28**(2), 461–482.

List of Author's Publications Used in Doctoral Thesis

Adam, L. and Bejda, P. (2017): 'Robust estimators based on generalization of trimmed mean', *Communications in Statistics-Simulation and Computation* (accepted).

Bejda, P. (2013): Generalization of geometric median, *in* H. Vojáčková, ed., 'Mathematical Methods in Economics 2013', Vol. 31, College of Polytechnics Jihlava, pp. 25–30.

Bejda, P. (2014): Generalization of geometric median, *in* J. Talašová, J. Stoklasa and T. Talášek, eds, 'Mathematical Methods in Economics 2014', Vol. 32, Palacky University in Olomouc, pp. 31–36.

Bejda, P. and Cipra, T. (2015): 'Exponential smoothing based on l-estimation', *Kybernetika* **51**(6), 973–993.