technische universität
dortmund

fakultät statistik

On nonparametric methods for robust jump-preserving smoothing and trend detection

Dissertation

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

Oliver Morell

August 2012

Betreuer und Gutachter: Prof. Dr. Roland Fried
Gutachter: Prof. Dr. Christine H. Müller
Kommissionsvorsitz: Prof. Dr. Jörg Rahnenführer
Tag der mündlichen Prüfung: 25. September 2012

Contents

Chapter 1

Introduction

1.1 Motivation

In terms of content this thesis deals with nonparametric smoothing and tests for detection of level shifts / jumps and trends.

Nonparametric smoothing is used to estimate an unknown regression function f based on some observations. The connection between f and the observations is disturbed by unobservable error terms. It is often advantageous to use nonparametric local estimation procedures rather than global parametric regression, as the nonparametric versions have a larger flexibility concerning the form of f, which has to be known in the parametric case, except a limited number of parameters. Nevertheless classical smoothing procedures based on means or kernel means can not handle discontinuities in f and are also not robust against outliers in the errors, meaning that largely deviating observations can strongly influence the estimation of f.

In nonparametric smoothing there are two primary targets, which are pursued:

(i) An overall good fit of the estimation of f.

(ii) Estimation / Detection of the jumps in f, particularly the number and location of the jump positions and if necessary the jump height.

Of course questions (i) and (ii) are interdependent. Incorrectly detected jumps due to a bad estimation of the true jump positions will lead to larger differences between true and estimated function values at design points close to this jump. Therefore an overall worse fit of f is the result and a distance measure like the Averaged Squared Error (ASE), which measures the quality of the overall fit by averaging the squared distances between true and estimated function values over all design points, will become larger then. On the other hand a large ASE-value can be due to a bad jump preservation of the smoother, but can also be caused by other properties of the true function or outliers in the errors, which make the used smoothing procedure inappropriate. To separate these two questions, Gijbels et al. (2007) distinguish between the direct and the indirect approach in jump-preserving smoothing.

In the direct approach the true function is estimated directly without any prior assumptions on number, location or height of the jumps. Each design point is considered as a potential discontinuity point and changes in the location level should be

incorporated automatically by the smoothing procedure. The main focus here lies on a good overall performance, e.g. measured by the ASE, what does not necessarily depend on a good jump detection. One way to smooth the true function is via local parametric fitting, e.g. to choose a local neighborhood around each design point and to use the corresponding observed values to estimate f.

The indirect approach includes the estimation of the jump positions as the first step. The resulting segments between the jumps are assumed to be continuous and can be estimated via a global procedure. The main focus of this approach lies on a good jump-detection-rule. The estimation of f in the continuous regions is secondary. One way to detect the jump locations is via test statistics which compare the location parameters between two samples. These tests are made at each design point, where the design point separates the two samples, respectively. The most likely design points of those, where the tests are rejected, are used as estimated jump locations.

For both approaches a proper choice of the smoothing parameter k is essential. For the direct approach, smoothing with a too large number of observations $2k+1$ in each window can lead to the loss of important details of f, e.g. non-preserved jumps, and so to a small variance, but a large bias of the estimation. On the other hand, a too small number of observations can cause an overfit to the observations and hence to a wiggly estimate, what brings a small bias, but large variance. For the indirect approach the parameter k delivers the sample size of the two samples for each test problem. For a too small chosen k not all true jumps may be detected due to a small power of the tests. For a too large chosen k only jumps in the middle of the design area can be detected, while the true jump locations are not necessarily located there. Thus variance and bias behave in a similar way as for the direct approach.

One way to choose k is via cross-validation (CV). Much work is done here for the direct approach, concerning optimality results of the classical L_2-CV (Haerdle and Marron, 1985; Haerdle, Hall and Marron, 1988), but also for robust CV-criteria (Leung et al., 1993; Wang and Scott, 1994; Zheng and Yang, 1998; Leung, 2005), which are needed in the presence of outliers in the noise terms, as L_2-CV performs poorly in this connection. However the theoretical results are always obtained under the assumption that the regression function is smooth, with a function f, which is two times differentiable or at least Hoelder-continuous. In particular this means also that f has no abrupt jumps. Further, as war as I know, practical comparisons of robust CV-criteria are also only made for the case of a smooth f.

For the indirect case only one cross-validated proposal is known to me, but it differs from our proposal. It uses the derivative of the Nadaraya-Watson-estimator (Nadaraya, 1964; Watson, 1964) as diagnostic function (Gijbels and Goderniaux, 2004a,b). A first rough estimation of the jump locations is given by those positions, where the derivative has a large value. These positions are improved then by a smooth least-squares estimation in a shorter sequence around the jump locations. However, it is only based on the L_2-CV-criterion and so for the indirect approach the behaviour of robust CV-criteria to my knowledge has not been investigated yet.

This thesis closes these two gaps in the literature, as it deals with a robust analysis of the direct approach, when additionally jumps in f occur, and with an analysis of the indirect approach, if additionally outliers in the errors are observed. Both prob-

lems are analysed via simulation studies as theoretical results for the cross-validated smoothers and tests are hard to determine. In a first step, piecewise constant functions are used to compare the performance of different cross-validated local constant smoothers and cross-validated test statistics.

The null hypothesis of the test problem from the indirect approach is the hypothesis of randomness, e.g. the observed values of both samples are one realisation of independent and identically distributed (i.i.d.) random variables. The alternative hypothesis includes a level shift between the two samples, i.e. both samples contain i.i.d. random variables, differing in a location parameter like the mean (or if not existing, e.g. the median). In contrast to that, in Chapter 4 the alternative hypothesis of a monotone trend is analysed. Here the random variables are still independent, but at least two of them are not identically distributed, as they differ in their location parameter. Furthermore, all location parameters are ordered by the time points, i.e. they are an increasing or decreasing sequence, building up a monotone trend.

Diersen and Trenkler (1996) reinvestigate tests based on records and Diersen and Trenkler (2001) apply a weighted and splitted form of these tests on a time series with seasonal effects. Splitting the observed time series sequence enlarges the robustness of the record tests against seasonality. The last part of the thesis deals with this fact and extends the idea of splitting the time series to several linear and nonlinear rank statistics. Beside the robustness against seasonality, the robustness against autocorrelation is examined. The procedures are further applied to two climate time series from the gauging station in Potsdam.

1.2 Outline

This thesis has two parts. The first part deal with nonparametric smoothing, in Chapter 2 with the direct and in Chapter 3 with the indirect approach. The second part in Chapter 4 closes with trend tests for seasonal time series. All Chapters in this thesis can be read individually, with identical notations in Chapter 2 and 3.

Chapter 2 investigates the direct approach for nonparametric smoothing, especially the performance of different robust CV-criteria, like median-CV (Zheng and Yang, 1998), M-CV (Leung et al., 1993; Leung, 2005; Bianco and Boente, 2006) with the Huber- and the Tukey-criterion, least trimmed squares-CV and a proposal by Bianco and Boente (2006), which considers the sum of a robustified estimation of the bias- and the variance-component of a cross-validated smoother. In addition to mean- and median-smoothers, some methods from signal processing and the commonly applied standard Lowess are taken for comparison.

The cross-validated smoothers are compared in situations with both piecewise constant and sine-functions under variation of design-parameters like percentage and magnitude of outliers, number and height of jumps and the sample size. The main results of Chapter 2 are the advantage of robust CV-criteria like Tukey- and Boente-CV in the presence of outliers and of M-CV-criteria, like the L_1, Huber- and Tukey-criterion in the presence of jumps. Overall, Tukey-CV seems to be an adequate choice of a CV-criterion, if outliers and jumps are present. Furthermore, L_1-CV, which is

often cited as a robust CV-criterion (Wang and Scott, 1994; Lee and Cox, 2010) performs poorly in the presence of large outliers, due to an unbounded loss function. The same is true for the smoother Lowess, which loses its robustness for a large number of outliers, independently of the used CV-criterion. This Chapter is accepted for publication under the name "On robust cross-validation for nonparametric smoothing".

In Chapter 3 the indirect approach is considered. Several jump detection rules are constructed under variation of a test statistic, a selection criterion for the level of significance, a selection rule for the jump candidates and a CV-criterion. Beside the classical t-test and linear rank tests, the comparative study also includes two sample test statistics based on trimmed means, medians and Hodges-Lehmann-estimators. For choosing the level of significance, multiple comparisons with the rules of Bonferroni and Bonferroni-Holm are considered as well as a fixed level for all test problems for all chosen sample sizes k. As one jump often leads to rejection of the Null for more than one design point close to the jump location, the selection rules of Wu and Chu (1993) and Qiu and Yandell (1998) are modified and compared, too.

The main results of Chapter 3 are a better performance of a multiple chosen level of significance, compared to a fixed one. The method of Bonferroni performs slightly better than the one of Bonferroni-Holm. Furthermore the procedure of Qiu and Yandell delivers closer estimations of the jump locations than the method of Wu and Chu for most of the tests. A small modification of Qiu and Yandell's original selection rule leads to an improvement here. Also in situations with outliers, robust CV-criteria are again superior to the classical L_2-CV-criterion. The Tukey-CV is again an adequate choice. From the test statistics, the two-sample-Hodges-Lehmann-test and the median-test perform best in the presence of large outliers. The content in this Chapter is unpublished yet, but considered for possible future publication.

Chapter 4 deals with a comparison of nonparametric tests based on records and ranks for trend detection in seasonal time series. Please note that the notation in this Chapter is slightly changed, due to a somewhat different data situation. The power of these nonparametric tests is compared for time series with four seasons and different sample sizes for linear, convex and concave trends. Furthermore, the detection rate of the tests is determined for observations, which are positive autocorrelated. For all tests beside the unsplitted version, different multiples of the number of seasons are used as splitting factors, respectively.

The main findings here are: While an increasing splitting factor increases the power of a record test, the power decreases for all reliable rank tests. Thereby the best rank tests based on the rank correlation coefficients of Spearman (Spearman, 1904) and Kendall (Kendall, 1938) perform better than any record test, even if the record test is splitted with a large splitting factor. However, the detection rates of the tests, if positive autocorrelated observations instead of a monotone trend are existent, deliver a smaller sensitivity against positive autocorrelation for the record tests than for the rank tests. This Chapter is published as:

Morell, O. and Fried, R. (2009): On nonparametric tests for trend detection in seasonal time series. In: Schipp, B., Krämer, W. (Eds): *Statistical Inference, Econometric Analysis and Matrix Algebra, Festschrift in Honour of Götz Trenkler*, Physica, Heidelberg, 19–40.178.

1.3 Outlook

The good performance of robust CV-criteria in case of local constant smoothers in the direct approach and two sample test statistics in the indirect approach leads to a couple of questions, which should be analysed in future work.

At this point only cross-validated local constant smoothers have been analysed for the direct approach. The use of local linear smoothers could bring advantages for functions with large slope or changes in the slope. Section 2.4.5 shows the advantage of Lowess as a local linear procedure, as long as the amplitude of the sine-function has a larger impact than the jump height. A similar performance can be expected for other local linear smoothing methods, while some robust regression methods could bring a better jump preservation, if they are used locally. A comparison of local linear smoothers with a smoothing parameter chosen by different robust CV-criteria is an interesting topic for future work. Beside classical L_2- and L_1-regression, robust regression methods like the repeated median (Siegel, 1982), least trimmed squares regression (Rousseeuw, 1984) or least quartile difference regression (Croux et al., 1994) can be used. All these linear regression methods can also be used for the indirect approach. Instead of comparing an estimated constant location measure of two samples, a comparison of the intercepts of the estimated straight lines of both samples can bring information about a jump within a curved function. This has been done for classical regression methods, but not for the above mentioned robust regression procedures. The problem of finding critical values can possibly be solved by the use of the permutation principle (see also Section 3.2.1), but will be computationally expensive.

The computational effort to compute the exact solution of an optimisation problem for a robust regression technique is generally a difficult task. This effort increases even more, when leave-one-out-CV is used for the selection of the smoothing parameter k, as a robust estimation is needed for each possible k and all other design points, when each design point is left out for one time. Algorithms, which deliver an updated estimation can help here (see Bernholt and Fried 2003 for the repeated median), as well as evolutionary computation (see Morell et al. 2008 for the least trimmed squares and Nunkesser and Morell 2010 for the least quartile difference regression). Further improvements to compute the CV-residuals for cross-validated smoothers and cross-validated test procedures are still needed and also an issue of future work.

Till now a global smoothing parameter k is chosen. Each locally chosen data window includes the same number of observations. A larger flexibility of the smoothing procedures can be achieved by a local adaptive smoothing parameter, especially if the true function has changes in the slope. This allows different choices of the smoothing parameter around different design points. Proposals for adaptive bandwidth-selectors like local cross-validation (Fan et al., 1996) are based on the L_2-norm and a comparison of these methods with a robust criterion can be promising.

The multivariate setting has also not been treated yet. Choosing a multidimensional smoothing parameter requires a reasonable preselection of possible tuples of window widths for each direction, to avoid a too large computational amount. Beside the robust CV-criteria proposed here, existing approaches for multivariate selection

of a window width (Kerkyacharian et al., 2001; Lafferty and Wasserman, 2008; Zhang et al., 2009) can be robustified. Again the computation of the robust smoothers and tests increases for a increasing dimension, what shows again the need of updating and evolutionary algorithms.

CV is not only used for nonparametric smoothing, but also in many other fields of statistical research. One example are time series, where alternative L_2-proposals to the classical leave-one-out-CV have been introduced (Francisco-Fernandez and Vilar-Fernandez, 2005). The use of robust CV-criteria instead of the L_2-criterion will give an improvement for choosing the smoothing parameter in robust time series analysis and can be a part of future work, too.

Chapter 2

On robust nonparametric smoothing: The direct approach

2.1 Introduction

We consider a regression model

$$Y_i = f(x_i) + E_i,\ i = 1, \ldots, n\ , \tag{2.1}$$

where f is an unknown piecewise continuous function, $x_1, \ldots, x_n$ are values of a covariate generated from a random design $X_1, \ldots, X_n$, $E_1, \ldots, E_n$ are i.i.d. errors possibly contaminated by some outliers, and $Y_1, \ldots, Y_n$ are observations of a response variable measured at $x_1, \ldots, x_n$, with realisation $y_1, \ldots, y_n$. For simplicity of the exposition we assume the data to be ordered according to the size of the x_i, $x_1 \leq x_2 \leq \ldots \leq x_n$.

Local parametric fitting allows us to estimate the unknown regression function f under weak assumptions, i.e. without the need of specifying a global functional form of f, which is known except for some unknown parameters. We concentrate on local constant smoothing here. Several such smoothers have been proposed, based on the idea to approximate f within suitably chosen local data windows by a constant.

The choice of the window width h is crucial for the performance of any local fitting method. If h is chosen small, the bias of the estimate becomes small and the variance large, leading to a wiggly estimate. If h is chosen large, the variance of the estimate gets smaller, but the bias increases. Important details of f can be lost then. A data-based approach to select the window width adaptively is cross-validation (CV). The following theoretical results are obtained under different assumptions, which always include f to be two times differentiable or at least Hoelder-continuous.

The basic idea of the commonly used leave-one-out-L_2-CV is to choose the window width h as the value that minimises the average squared distance between the true observations $y_1, \ldots, y_n$ and the leave-one-out-estimates $\widehat{f}_{-1;h}(x_1), \ldots, \widehat{f}_{-n;h}(x_n)$, see Section 2.3 for details. Haerdle and Marron (1985) proved the asymptotic optimality of L_2-CV for linear kernel regression estimators with respect to goodness-of-fit-measures based on the L_2-norm, like the Integrated Squared Error (ISE), the

Conditional Mean Integrated Squared Error (CMISE) or the Averaged Squared Error (ASE). Asymptotic optimality with respect to a distance measure Δ is defined as

$$\lim_{n\to\infty}\left(\frac{\Delta\left(\widehat{f}_{\widehat{h}_{\mathrm{CV}}}(x_i), f(x_i)\right)_{i=1,\ldots,n}}{\Delta\left(\widehat{f}_{\widehat{h}_{\Delta}}(x_i), f(x_i)\right)_{i=1,\ldots,n}}\right) = 1, \tag{2.2}$$

with probability one, see Shibata (1981). In this formula $\widehat{h}_{\mathrm{CV}}$ is the L_2-cross-validated and $\widehat{h}_{\Delta}$ is the Δ-optimal window width. For nonlinear kernel M-smoothers, Haerdle (1984) showed the convergence in probability of $\widehat{h}_{\mathrm{CV}}$ against the ASE-optimal window width under the assumption of finite second moments. Although $\widehat{h}_{\mathrm{CV}}$ and the ASE-optimal window width $\widehat{h}_{\mathrm{ASE}}$ are asymptotically the same, Haerdle, Hall and Marron (1988) showed that the relative difference $(\widehat{h}_{\mathrm{CV}} - \widehat{h}_{\mathrm{ASE}})/\widehat{h}_{\mathrm{ASE}}$ has only a slow rate of convergence of $n^{-1/10}$ and so for different samples from the same model, $\widehat{h}_{\mathrm{CV}}$ can lead to rather different window widths and also to quite different estimations of f than $\widehat{h}_{\mathrm{ASE}}$. Nevertheless it seems to be the best possible rate, as the relative difference between the optimal window widths for the ASE and the mean of the ASE converge with the same rate (Haerdle, Hall and Marron, 1988).

Both smoother and CV-criterion should be chosen robustly if outliers occur (Leung et al., 1993). The commonly used L_2-CV is not robust, so alternatives like L_1-CV (Yang and Zheng, 1992; Wang and Scott, 1994), M-CV (Leung et al., 1993; Cantoni and Ronchetti, 2001; Leung, 2005) and median-CV (Zheng and Yang, 1998) have been proposed. We also consider the least trimmed squares criterion of Rousseeuw (1984), applied by Serneels et al. (2005), and a modified version of a CV-proposal of Bianco and Boente (2006) and Boente and Rodriguez (2008). For kernel M-smoothers, the window width selected via M-CV is asymptotically equivalent to $\widehat{h}_{\mathrm{ASE}}$ (Leung, 2005).

The optimal window width $\widehat{h}_{\Delta}$ for a goodness-of-fit-measure Δ based on squared distances depends for linear (e.g. Haerdle (2002), pp. 30) and nonlinear smoothers (Wang and Scott, 1994; Zheng and Yang, 1998; Leung, 2005) on the second derivative of the true function f. Asymptotic optimality as in (2.2) and even consistency is unclear, if f has edges or abrupt jumps, what are effects of discontinuities in the first derivative or f itself.

In practice the assumption of a smooth regression function is often not appropriate, because of jumps or edges. As there is a lack of theory and experimental studies in situations with discontinuities, we check via simulations which cross-validated smoothers yield the best results then. We compare the different CV-criteria in situations with jumps in the regression function f and outliers in the errors $E_1 \ldots, E_n$. We focus on classical local constant smoothers based on means or medians and methods from signal processing, like double window and linear hybrid smoothing methods.

Section 2.2 reviews local constant smoothers. Section 2.3 introduces different CV-criteria. Section 2.4 describes the results of a simulation study, comments consistency of the estimators and includes some applications to real data. Section 2.5 concludes.

2.2 Local constant smoothers

We distinguish between moving window (MW) and nearest neighbour (NN) smoothers. A MW-smoother estimates f at each point x_i by a location estimator of the observations at the $2k+1$ design points $x_{i-k} \ldots, x_{i+k}$ centered at x_i, if available. For the first and the last k design points we take the estimations based on the first and the last k observations, respectively. Using the sample mean we get the moving average

$$\Xi_1(x_i) = \frac{1}{2k+1} \sum_{j=-k}^{k} y_{i+j} \ . \tag{2.3}$$

Let $x_{i,(j)}$ be the j-th nearest neighbour of x_i, $\left|x_{i,(1)} - x_i\right| \leq \left|x_{i,(2)} - x_i\right| \leq \ldots \leq \left|x_{i,(n)} - x_i\right|$, and $y_{i,(j)}$ the value observed at $x_{i,(j)}$, for $j = 1, \ldots, n$. Then a κ-NN-smoother is a location measure of the observations $y_{i,(1)}, \ldots, y_{i,(\kappa)}$. We choose $\kappa = 2k+1$ to use the same number of observations for MW- and NN-smoothers in each window. The NN-mean is then defined as

$$\Xi_2(x_i) = \frac{1}{2k+1} \sum_{j=1}^{2k+1} y_{i,(j)}. \tag{2.4}$$

A MW- and a NN-smoother with the same location measure give identical results at $x_{k+1}, \ldots, x_{n-k}$ in case of an equidistant fixed design, but are different in general, since the NN are not necessarily distributed equally to the left and the right of x_i.

An advantage of mean-smoothers is their high efficiency in case of normal errors. However, a single outlier affects the estimation and can make it completely meaningless locally. The robustness of an estimate against outliers can be measured by the finite sample breakdown point (Donoho and Huber, 1983). It corresponds to the minimal fraction of modifications in a sample which can drive the estimate to the boundaries of the parameter space. In case of a sample of size $2k+1$ it is $1/(2k+1)$ for the sample mean, meaning that a single outlier can cause a spike of any size in the estimate $\Xi_1(x_i)$ or $\Xi_2(x_i)$. Moreover, mean smoothers smear jumps, leading to strongly biased and even inconsistent estimates there, since the estimates average observations before and after the jump-location.

Median smoothers improve upon both these shortcomings. The MW-median

$$\Xi_{3,a}(x_i) = \mathrm{Med}(y_{i-k}, \ldots, y_{i+k}) \ , \tag{2.5}$$

which is also called running median in the literature, and the NN-median

$$\Xi_{4,a}(x_i) = \mathrm{Med}(y_{i,(1)}, \ldots, y_{i,(2k+1)}) \tag{2.6}$$

both offer a finite sample breakdown point of $(k+1)/(2k+1)$ within each window, which is optimal within the class of all location-equivariant estimators. Moreover, jumps between two constant parts of the function are preserved if there are at least $(k+1)$ observations for each part available and the smoother uses the same number of observations left and right of the jump. In a random design this is guaranteed

for the MW-median, but not for the NN-version. Under Gaussian noise, the sample median offers an asymptotic efficiency of only 63.7% relatively to the sample mean.

Other MW- and NN-smoothers based on robust location measures like the Huber- and Tukey-M-estimator (Maronna et al., 2006, pp. 22–31), the 12.5%- and 25%-trimmed mean, the 50%- and 75%-least trimmed squares (LTS) estimator (Rousseeuw, 1984) or the Hodges-Lehmann-estimator (Hodges and Lehmann, 1963) can be defined analogously. In a preliminary simulation study only the MW-version of the 50%-LTS-smoother showed both adequate jump-preservation and robustness against outliers, but performed worse compared to the MW-median. To keep the simulation study manageable none of these smoothers will be considered in the following.

Another approach from signal processing for local constant function fitting are linear median hybrid (LMH) filters (Heinonen and Neuvo, 1987). The outputs of m linear subfilters $H_1, \ldots, H_m$ are calculated for each x_i and their median is taken to estimate $f(x_i)$. As proposed by Heinonen and Neuvo, we use $m = 3$ and define

$$\begin{aligned} \Xi_5(x_i) &= \text{Med}(H_1(y_i), H_2(y_i), H_3(y_i)), \quad \text{with} \\ H_1(y_i) = \frac{1}{k}\sum_{j=1}^{k} y_{i-j},\ H_2(y_i) &= y_i,\ H_3(y_i) = \frac{1}{k}\sum_{j=1}^{k} y_{i+j}\ . \end{aligned} \tag{2.7}$$

$H_1(y_i)$ and $H_3(y_i)$ take the average of the k observations left and right of the current design point x_i, respectively, whereas the output of $H_2(y_i)$ is y_i itself to improve the preservation of jumps. To increase the robustness against outliers, Fried et al. (2006) use the median instead of the average in the subfilters and derive median median hybrid (MMH) smoothers,

$$\begin{aligned} \Xi_6(x_i) &= \text{Med}(M_1(y_i), M_2(y_i), M_3(y_i)), \quad \text{with} \quad M_2(y_i) = y_i, \\ M_1(y_i) &= \text{Med}(y_{i-k}, \ldots, y_{i-1}),\ M_3(y_i) = \text{Med}(y_{i+1}, \ldots, y_{i+k})\ . \end{aligned} \tag{2.8}$$

Another method from signal processing based on MW is the double window modified trimmed mean (DWMTM), see Lee and Kassam (1985). Defining a trimming factor $\varpi \in (0, 0.5)$, a ϖ-trimmed mean is an average of the observations, with the $100\varpi\%$ smallest and the $100\varpi\%$ largest values being disregarded. Trimmed means achieve larger efficiency under normal noise than the sample median, which corresponds to the limiting case $\varpi = 0.5$, but do not preserve jumps exactly. Therefore a procedure with an adaptive, data-based choice of ϖ, like the DWMTM, is preferable.

The DWMTM, denoted by Ξ_7, uses two windows with smoothing parameters k and l, respectively, with $l \leq k$. The median $\tilde{y}_i$ and the median absolute deviation from the median (MAD) $\widehat{s}_M$ as a robust measure of location and variability, respectively, are calculated from the possibly smaller inner window $y_{i-l}, \ldots, y_{i+l}$. Then all observations $z \in \{y_{i-k}, \ldots, y_{i+k}\}$ with $|z - \tilde{y}_i| > \delta \cdot \widehat{s}_M$ are trimmed and the remaining values are averaged. Here, δ is a predefined constant regulating the amount of trimming. We choose $\delta = 2$, as proposed by Lee and Kassam, and fix $l = \lfloor k/2 \rfloor$ for the reason of simplicity. Working with a short inner window gives little biased initial estimates of

location and variability. The subsequent averaging step with an adaptively chosen trimming constant reduces the variability of the final estimate.

We also consider a robust version of locally weighted regression (Cleveland, 1979), called Lowess and denoted by Ξ_8 here, since it is a commonly applied standard. Let W_3 be the bisquare function and W_2 be the tricube function, with

$$W_\beta(x) \;=\; \begin{cases} (1-|x|^\beta)^\beta & , \text{ if } |x|<1 \\ 0 & , \text{ if } |x|\geq 1 \end{cases} , \; \beta = 2,3 \tag{2.9}$$

and d_{ik} be the absolute distance between x_i and its $(2k+1)$-th NN $x_{i,(2k+1)}$. Using the weights $w_j(x_i) = W_3(\frac{x_j - x_i}{d_{ik}}), j = 1,\ldots,n$, for each data point (x_j, y_j), a locally weighted regression is done in an initial step. By construction these weights are decreasing for an increasing distance of x_i to x_j and zero, if x_j is none of the first $2k$ NN of x_i. The obtained residuals $\widehat{e_1},\ldots,\widehat{e_n}$ are used to derive new robustness weights $\delta_1,\ldots,\delta_n$, with $\delta_j = W_2\left(\frac{\widehat{e_j}}{6\widehat{e}_{Med}}\right)$, where $\widehat{e}_{Med}$ is the median absolute residual. Then a new locally weighted regression, using the robust localized weight $\delta_j \cdot w_j(x_i)$ is calculated for each (x_j, y_j). Observations with large residuals compared to $\widehat{e}_{Med}$ get a small weight or are trimmed completely. This step can be repeated, calculating another robust locally weighted regression with weights based on the previous step. Cleveland states that two iterations are typically enough. We use the R-function Lowess (R Development Core Team, 2011) for the computations. Lowess is often called a robust smoothing method, but as it uses a non-robust weighted least squares regression as initial estimation, its robustness against outliers is questionable (Maechler, 1989).

2.3 Robust cross-validation

The performance of every procedure described in Section 2.2 depends on its smoothing parameter k, which delivers the number of observations $2k+1$ within each window and the window width h. A way to choose k adaptively, i.e. based on the data, is cross-validation (CV). For a given k let $\widehat{f_k}(x_i)$ be the estimate of f at x_i derived by one of the smoothers $\Xi_1,\ldots,\Xi_8$. Then $\widehat{f}_{-i;k}(x_i)$ is the estimate of f, when the point (x_i, y_i) itself is left out and not used for the estimation. For the moving window techniques we base the estimation $\widehat{f}_{-i;k}(x_i)$ on the $2k+1$ points in $\{x_{i-1-k},\ldots,x_{i-1},x_{i+1},\ldots,x_{i+k}\}$. We denote the cross-validated residuals by $\widehat{e}_{i;k} = y_i - \widehat{f}_{-i;k}(x_i), i = 1,\ldots,n$.

A common criterion for the choice of k is the traditional least squares CV, or briefly L_2-CV. It minimises the averaged sum of squared residuals as a function of k:

$$LSCV(k) \;=\; \arg\min_k \frac{1}{n}\sum_{i=1}^n \left(y_i - \widehat{f}_{-i;k}(x_i)\right)^2 \;=\; \arg\min_k \frac{1}{n}\sum_{i=1}^n \widehat{e}_{i;k}^{\,2} \,. \tag{2.10}$$

Note that the expectation of $LSCV(k)$ becomes infinite for each k if the second moment of the error distribution does not exist (Leung et al., 1993). So a single outlier in moderate samples can already lead to a nearly constant $LSCV(k)$ due to

its large squared error for all values of k (Wang and Scott, 1994). This makes L_2-CV practically useless in the presence of outliers, see also Section 2.4.

An alternative is to use absolute instead of squared distances. This leads to least absolute deviations CV, or briefly L_1-CV:

$$LADCV(k) \;=\; \arg\min_k \frac{1}{n}\sum_{i=1}^{n} |y_i - \widehat{f}_{-i;k}(x_i)| \;=\; \arg\min_k \frac{1}{n}\sum_{i=1}^{n} |\widehat{e}_{i;k}|. \tag{2.11}$$

Wang and Scott (1994) introduced locally weighted L_1-regression, combined with L_1-CV as a robust alternative to L_2-CV, to estimate f. Their simulations indicated that L_1-CV works better than L_2-CV if the error distribution is heavy-tailed. Pointwise consistency was also shown for the L_1-regression smoother with a theoretically optimal k in case of two times continuous differentiable regression functions.

Yang and Zheng (1992) considered L_1-CV for NN-median smoothing and proved weak consistency of this cross-validated smoother if f is Hoelder-continuous and the first moment of the errors exists. If the latter assumption is not fulfilled, the expectation of $LADCV(k)$ becomes infinite, meaning that L_1-CV may have problems with outliers. Therefore Zheng and Yang (1998) introduced median-CV

$$MEDCV(k) \;=\; \arg\min_k \operatorname{Med}\left(|\widehat{e}_{1;k}|, |\widehat{e}_{2;k}|, \ldots, |\widehat{e}_{n;k}|\right), \tag{2.12}$$

in combination with NN-median smoothing, as a robust alternative to L_1- and L_2-CV. In their work they show the uniform strong consistency of median-CV under Hoelder-continuity of f and compare it via simulations with L_1- and L_2-CV, but only for one data situation with outliers, where median-CV delivered better results than the two competitors. A possible disadvantage of median-CV is that a lot of information gets lost for the determination of k since only the median of the absolute residuals is used. Therefore we also consider other robust measures different from the median.

One alternative is M-CV, see Leung et al. (1993), Leung (2005) and Lee and Cox (2010). Bianco and Boente (2006) considered

$$M_\rho CV(k) \;=\; \arg\min_k \frac{1}{n}\,\widehat{\sigma}_k^2 \sum_{i=1}^{n} \rho\left(\frac{\widehat{e}_{i;k}}{\widehat{\sigma}_k}\right) . \tag{2.13}$$

Note that the robust error variance estimate $\widehat{\sigma}_k^2$ depends on the cross-validated residuals and hence on the smoothing parameter k. We consider two different ρ-functions for M-CV. Denoting the indicator function of a subset $A \subset \mathbb{R}$ by $\mathbb{1}_A$, we get a monotone M-criterion using the Huber-ρ-function

$$\rho_H(z) = z^2 \mathbb{1}_{[0,\ell_H]}(|z|) + \left(2\ell_H|z| - \ell_H^2\right) \mathbb{1}_{(\ell_H,\infty)}(|z|) \tag{2.14}$$

and a redescending M-criterion using the Tukey-ρ-function

$$\rho_T(z) = \left(1 - \left(1 - \left(\frac{z}{\ell_T}\right)^2\right)^3\right) \mathbb{1}_{[0,\ell_T]}(|z|) + \mathbb{1}_{(\ell_T,\infty)}(|z|) \tag{2.15}$$

to derive CV-criteria with possibly different properties. We choose $\ell_H = 1.345$ and $\ell_T = 4.685$ as tuning constants to achieve an efficiency of 95% for both criteria at the normal distribution in a location problem. We found the Q_n (Rousseeuw and Croux, 1993) estimator for σ_k to give better results than the MAD or the τ-scale estimator (Maronna and Zamar, 2002) for both ρ-functions. The Q_n corresponds roughly to the first quartile of the ordered absolute pairwise differences $|\widehat{e}_{i;k} - \widehat{e}_{j;k}|, i \neq j$.

For the Huber-ρ-function, one can also consider the following modified criterion

$$\arg\min_k \left(\frac{1}{n} \sum_{i=1}^{n} \widehat{e}_{i;k}^{\,2} \mathbb{1}_{[0,\ell_H]} \left(\frac{|\widehat{e}_{i;k}|}{\widehat{\sigma}_k} \right) + \frac{1}{n} \sum_{i=1}^{n} |\widehat{e}_{i;k}| \mathbb{1}_{(\ell_H,\infty)} \left(\frac{|\widehat{e}_{i;k}|}{\widehat{\sigma}_k} \right) \right), \tag{2.16}$$

but we found a larger robustness of the Huber-CV as in (2.13) in a preliminary study.

Serneels et al. (2005) used a robust CV-criterion based on least trimmed squares (LTS) for continuum regression. It is defined by

$$LTS_tCV(k) = \arg\min_k \frac{1}{\lfloor tn \rfloor} \sum_{j=1}^{\lfloor tn \rfloor} \widehat{e}_{(j);k}^{\,2} , \tag{2.17}$$

where $(1-t) \in (0, 0.5)$ is a trimming factor and $\widehat{e}_{(j);k}^{\,2}, j = 1, \ldots, n$ are the order statistics of the squared cross-validated residuals. The $\lfloor tn \rfloor$ smallest squared residuals are averaged and the others are trimmed. We consider the cases $t = 0.5$ to achieve a highly robust criterion and $t = 0.75$ to compromise between robustness and efficiency.

Another class of CV-criteria was introduced by Bianco and Boente (2006) and Boente and Rodriguez (2008) in the semiparametric regression case. Their criterion considers the sum of the squared bias and the variance of the cross-validated smoother. They estimate both components robustly, the squared bias via the squared median residual and the variance via some robust estimator $\widehat{\sigma}_k^2$. The authors recommend the τ-scale estimator, but we found in preliminary simulations again a better performance of the Q_n estimator. So we define a modified criterion as

$$BOECV(k) = \arg\min_k \operatorname{Med}^2 (\widehat{e}_{1;k}, \ldots, \widehat{e}_{n;k}) + Q_n^2 (\widehat{e}_{1;k}, \ldots, \widehat{e}_{n;k}) . \tag{2.18}$$

In most of the papers cited above, the proposed robust CV-criterion is only compared to the classical L_2-CV. For robust kernel smoothing, Leung (2005) compared L_1-CV with a Huber-criterion similar to (2.16), which has the same problems as L_1-CV if the first moment of the error distribution does not exist, due to its unboundedness. Bianco and Boente (2006) find their proposal to perform slightly better than M-CV (2.13) with the Huber- and the Tukey-ρ-function, respectively, for semi-parametric partly linear autoregression.

None of these papers examines situations with a discontinuous regression function f. Theoretical results are always obtained under the basic assumption that f is at least two times continuously differentiable or Hoelder-continuous. As there are no theoretical or practical comparisons between robust smoothers or CV-criteria for discontinuous regression functions available, we analyse the behaviour of the robust procedures, when this basic assumption is violated due to jumps in f.

2.4 Comparison of the cross-validated smoothers

In this Section we present a simulation study to compare the performance of the smoothers and CV-criteria for different data situations with outliers and jumps. For every situation each of the eight smoothers $\Xi_1, \ldots, \Xi_8$ from Section 2.2 is combined with one CV-criterion of each class (L_1-, L_2-, median-, Huber-, Tukey-, LTS- and Boente-CV). For the Huber- and Tukey-CV we use (2.13) with the Q_n scale estimator. Unless indicated otherwise, we use 75%-LTS-CV instead of 50%-LTS-CV due to its better efficiency. The Boente-CV will be computed as in (2.18). If any CV-criteria delivers its smallest criterion-value for more than one argument k, then the minimum of these arguments is taken as cross-validated smoothing parameter for estimating f.

2.4.1 Measuring the quality of a cross-validated smoother

We consider model (2.1) with $X_1, \ldots, X_n$ being drawn from an uniform random design, i.e. $X_1, \ldots, X_n$ being i.i.d. uniformly distributed on the interval $[0, 1]$. The noise terms $E_1, \ldots, E_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ with some outliers at positions chosen at random. For obtaining π percent outliers, $\max\{\lfloor n\pi \rfloor, 1\}$ of all n positions are drawn without replacement. At outlier positions, the value $\pm\gamma$ is added to the noise, with the same sign as the closest level shift to produce a more challenging situation. The function f is chosen as piecewise constant with m jumps of height s, whose positions $\xi_1, \ldots, \xi_m$ within the interval (0,1) are fixed. We choose γ and s depending on the error variance σ^2, here for $\sigma = 1$. We use a piecewise constant function since we assume the effects of jumps and outliers on the estimates to be more severe than a slight slope.

Let $\widehat{f}(x_i^\lambda)$ be the estimate of f for a given data set λ at its i-th design point x_i^λ, derived by one of the smoothers $\Xi_1, \ldots, \Xi_8$ combined with a CV-criterion of Section 2.3. The performance of $\widehat{f}$ for the λ-th data set can be measured by the Averaged Squared Error (ASE; see Haerdle 2002, pp. 90)

$$\widetilde{\Delta}_A^\lambda = \frac{1}{n} \sum_{i=1}^{n} \left(\widehat{f}(x_i^\lambda) - f(x_i^\lambda) \right)^2 . \tag{2.19}$$

The ASE averages the squared distances between true and estimated function values at all design points of one data set. We use the design points, because these are the only positions in the support, where we have information about the true function. Additional comparisons gave only small changes when evaluating the fit on an equally spaced grid.

If ν data sets are generated for a given data situation, the mean ASE-value (MASE)

$$\overline{\Delta}_A = \frac{1}{\nu} \sum_{\lambda=1}^{\nu} \widetilde{\Delta}_A^\lambda = \frac{1}{\nu n} \sum_{\lambda=1}^{\nu} \sum_{i=1}^{n} \left(\widehat{f}(x_i^\lambda) - f(x_i^\lambda) \right)^2 \tag{2.20}$$

can be calculated. Here $\widetilde{\Delta}_A^\lambda$ is the ASE-value and x_i^λ is the i-th design point for the λ-th data set, $\lambda = 1, \ldots, \nu$.

Instead of squared distances in (2.19) the average

$$\widetilde{\Delta}_B^\lambda = \frac{1}{n}\sum_{i=1}^{n} |\widehat{f}(x_i^\lambda) - f(x_i^\lambda)|. \tag{2.21}$$

and the median value

$$\widetilde{\Delta}_C^\lambda = \text{Med}(|\widehat{f}(x_1^\lambda) - f(x_1^\lambda)|, \ldots, |\widehat{f}(x_n^\lambda) - f(x_n^\lambda)|) \tag{2.22}$$

of the absolute distances between the true and the estimated function values can be used, denoted by the Averaged Absolute Error (AAE) and the Median Absolute Error (MAE), respectively. For the comparison of ν data sets, appropriate performance measures are the mean AAE-value (MAAE)

$$\overline{\Delta}_B = \frac{1}{\nu}\sum_{\lambda=1}^{\nu} \widetilde{\Delta}_B^\lambda = \frac{1}{\nu n}\sum_{\lambda=1}^{\nu}\sum_{i=1}^{n} |\widehat{f}(x_i^\lambda) - f(x_i^\lambda)| \tag{2.23}$$

and the median MAE (MMAE)

$$\overline{\Delta}_C = \text{Med}(\widetilde{\Delta}_C^1, \ldots, \widetilde{\Delta}_C^\nu) \tag{2.24}$$

For these two alternative performance measures larger distances, due to a bad estimation of f at only some design points or only for some data sets, are less relevant than for the MASE. Therefore the MAAE and the MMAE can favor other cross-validated smoothers than the MASE. We will use the MASE $\overline{\Delta}_A$ in the following and set $n = 200$ and $\nu = 1000$ for all data situations. In Section 2.4.4 we will present results for other sample sizes n and the MAAE and the MMAE, respectively.

It is also of interest to compare the performances for p different data situations jointly. We have $q = 56$ estimators combining all smoothers with all CV-criteria. In order to simplify this evaluation, we define a summary measure to compare their relative performances. For the η-th data situation, $\eta = 1, \ldots, p$, we consider the relative loss

$$\Lambda_\eta^\upsilon = \frac{\overline{\Delta}_{A;\eta}^\upsilon - \overline{\Delta}_{A;\eta}^\star}{\overline{\Delta}_{A;\eta}^\star}, \tag{2.25}$$

with $\overline{\Delta}_{A;\eta}^\upsilon$ as the MASE of estimator $\widehat{f}_\upsilon, \upsilon = 1, \ldots, q$, for data situation η and

$$\overline{\Delta}_{A;\eta}^\star = \min\left(\overline{\Delta}_{A;\eta}^1, \ldots, \overline{\Delta}_{A;\eta}^q\right) \tag{2.26}$$

the minimal MASE-value of all estimators for data situation η. So Λ_η^υ is the relative loss in the MASE-value due to not using the best estimator, relatively to $\overline{\Delta}_{A;\eta}^\star$. Values of Λ_η^υ close to zero indicate that $\widehat{f}_\upsilon$ performs almost as well as the best estimator for situation η. We use the mean relative loss (MRL) in the MASE-value

$$\overline{\Lambda}^\upsilon = \frac{1}{p}\sum_{\eta=1}^{p} \Lambda_\eta^\upsilon \tag{2.27}$$

as global performance measure in the comparison for the included data situations.

2.4.2 Robust smoothing without jumps

We start with situations, where problems are mainly caused by outliers and not by jumps. We draw $n = 200$ observations (x_i, y_i) with one small jump of height 1 at the position $\xi_1 = 0.4$ and vary the percentages $\pi \in \{0.01, 0.05, 0.15\}$ and magnitudes $\gamma \in \{3, 6, 12\}$ of the outliers. The MRL $\overline{\Lambda}^v$ from (2.27) based on these $p = 9$ data situations is illustrated in Fig. 2.1 (left). We use a logarithmic scale for the ordinate since we want to visualise differences among the better estimators, accepting less visibility of the differences among the worse estimators. The results for the mean and the LMH smoothers are not shown, because of their bad performance due to the lack of robustness against outliers.

Lowess performs best, followed by DWMTM. Tukey-CV gives the best results for all smoothers except for the MMH, followed by Boente- and Huber-CV. A closer look at the loss for the different situations is given for Lowess in Table 2.1. The order of the CV-criteria is similar for the other robust smoothers. The bad performance of L_1-CV is due to the situation with many large outliers, indicating its smaller robustness compared to the other robust CV-criteria. This leads us to look at situations with higher values of γ and π.

Looking at larger outliers, we take again $\pi \in \{0.01, 0.05, 0.15\}$ and vary γ in $\{24, 48, 96, 192\}$. Fig. 2.1 (right) shows the MRL for all robust smoothers based on these $p = 12$ data situations. Generally, Tukey-, Boente- and LTS-CV lead to the best results for all robust smoothers. This can also be seen in Fig. 2.2, where the MASE-values for an increasing outlier magnitude are shown for the MW-median and

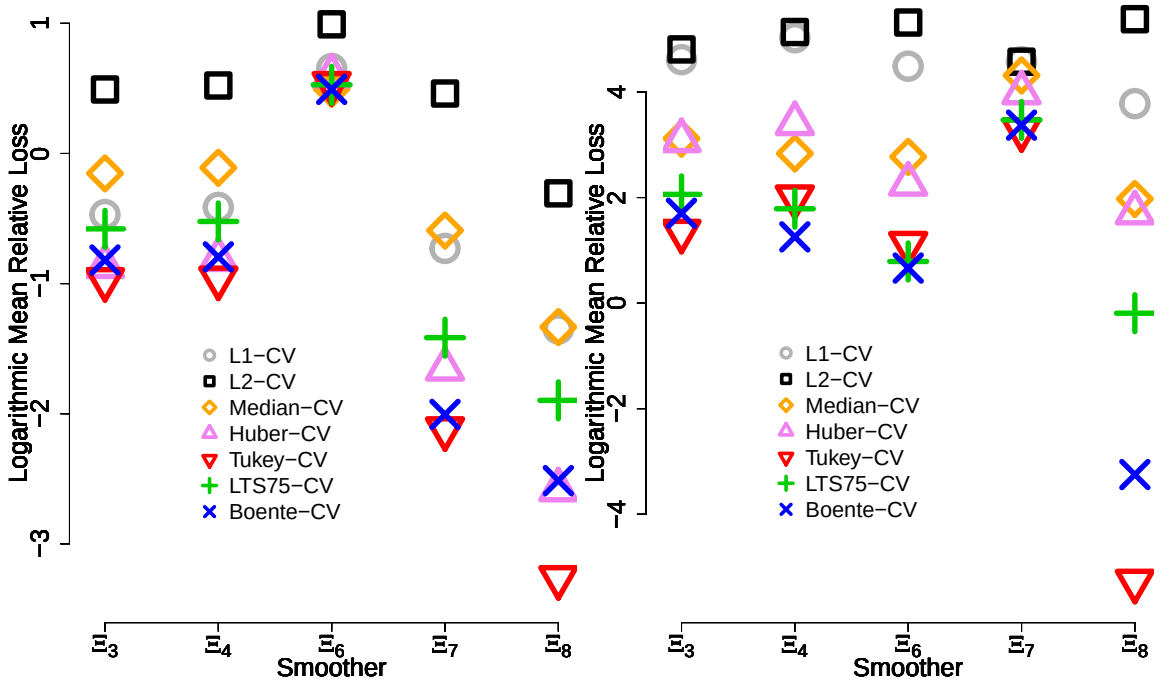

Figure 2.1: MRL for moderate percentages of moderate (left) and large (right) outliers. Lowess (with Tukey-CV) delivers the smallest relative loss.

Table 2.1: Relative loss of Lowess with different CVs for situations with moderately large outliers.

situation	Λ_η^v for Lowess with different CV-criteria						
(π, γ)	$LADCV$	$LSCV$	$MEDCV$	$M_{Hub}CV$	$M_{Tuk}CV$	$LTS_{75}CV$	$BOECV$
(0.01,3)	0.095	0.030	0.320	0.045	0.047	0.229	0.126
(0.01,6)	0.065	0.022	0.261	0.020	0.025	0.164	0.077
(0.01,12)	0.033	0.136	0.273	0.000	0.001	0.184	0.067
(0.05,3)	0.078	0.040	0.215	0.028	0.035	0.172	0.104
(0.05,6)	0.032	0.111	0.236	0.000	0.000	0.148	0.049
(0.05,12)	0.087	0.556	0.271	0.054	0.000	0.162	0.068
(0.15,3)	0.196	0.216	0.345	0.197	0.199	0.218	0.209
(0.15,6)	0.157	1.004	0.165	0.044	0.037	0.031	0.004
(0.15,12)	1.595	4.511	0.287	0.291	0.000	0.042	0.024

Lowess. Fig. 2.2 also points out that Lowess is only robust against large outliers, if an appropriate CV-criterion is used. L_1-CV, introduced by Lee and Cox (2010) as a robust alternative to L_2-CV for Lowess, is not appropriate here, as it performs poorly for larger outliers. This confirms the results of Zheng and Yang (1998) that a robust CV-criterion is preferable if the data are contaminated by large outliers, but the median-CV proposed by them is also outperformed by all other robust CV-criteria.

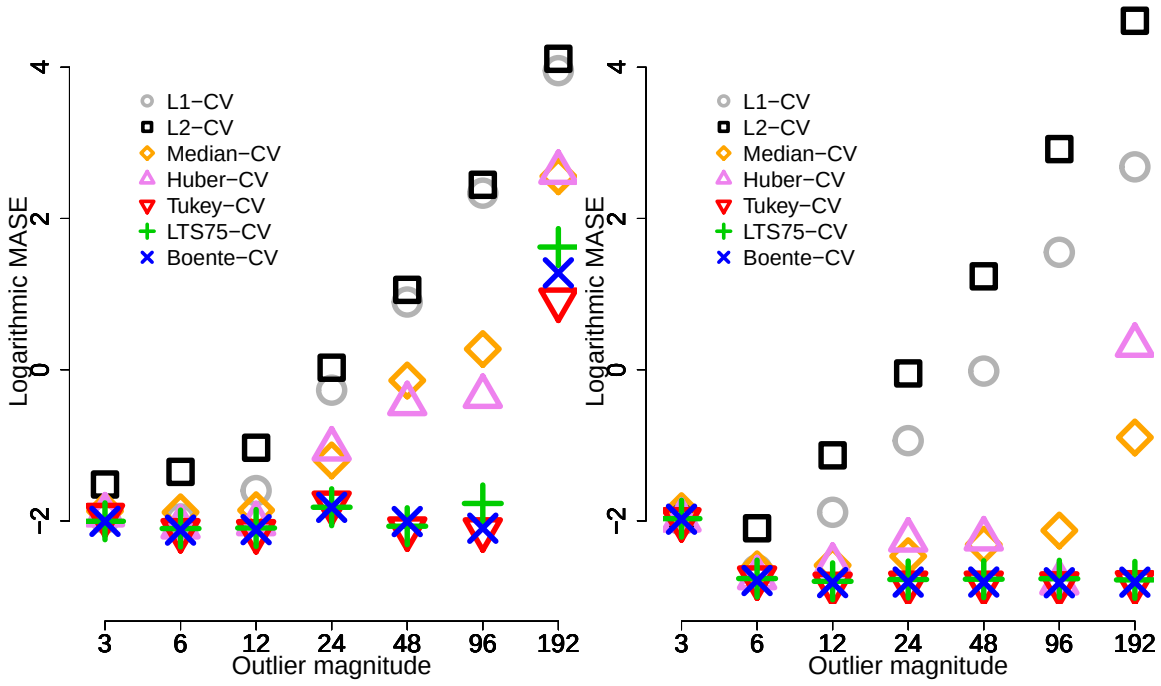

Figure 2.2: MASE-values for an increasing magnitude of outliers for the MW-median (left) and Lowess (right). L_2- and L_1-CV perform worst for an increasing outlier magnitude.

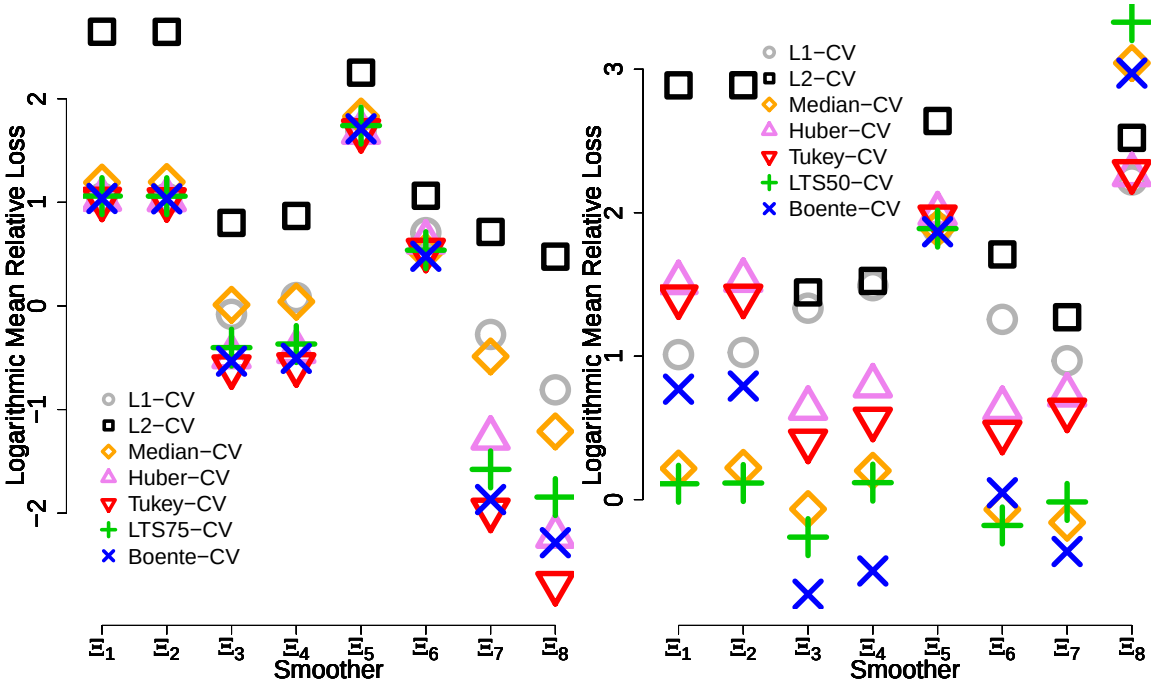

Figure 2.3: MRL for situations with moderate (left) and large (right) percentages of outliers. Boente-CV gives good results for moderate and large outlier percentages.

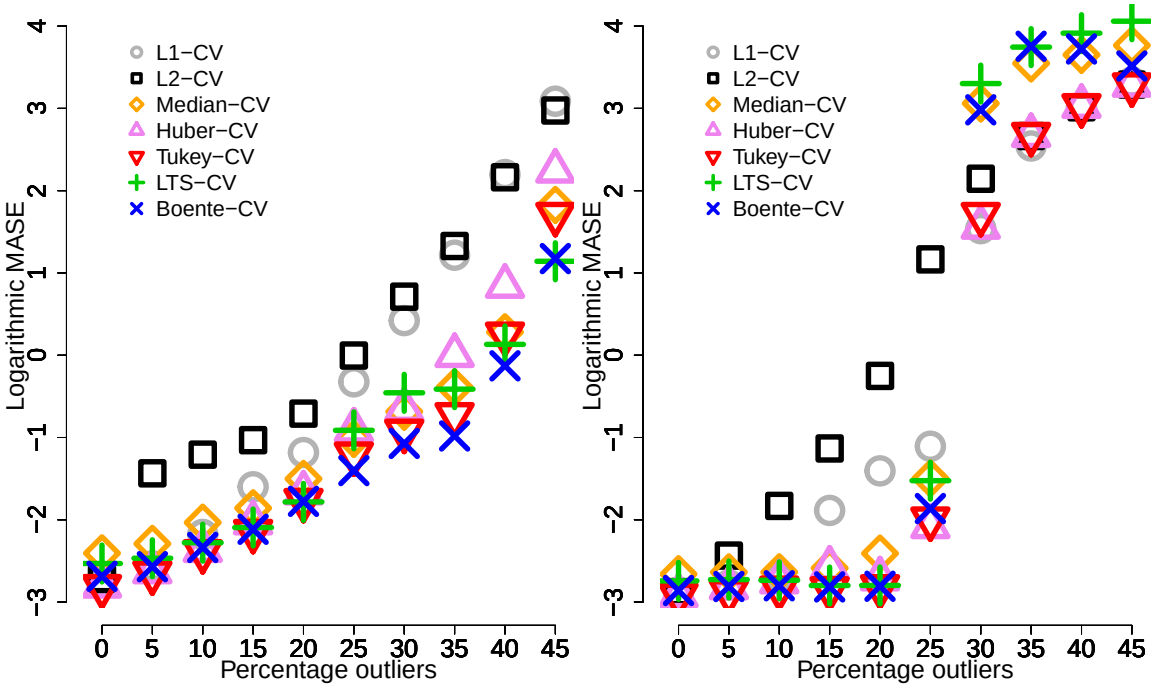

Figure 2.4: MASE-values for an increasing percentage of outliers for the MW-median (left) and Lowess (right). Lowess loses its robustness for large outlier percentages.

Next we increase the numbers of outliers. We vary π in $\{0, 0.05, \ldots, 0.45\}$ and use moderate outlier magnitudes $\gamma \in \{3, 6, 12\}$. Fig. 2.3 shows the MRL-values for situations with moderate ($\pi \leq 0.2$) and large ($\pi \geq 0.25$) percentages of contamination. Note that we have replaced 75%-LTS-CV for the larger percentages $\pi \geq 0.25$ with the more robust version, the 50%-LTS-CV.

Lowess loses its robustness for $\pi \geq 0.3$, and the median-smoothers become preferable. From the CV-criteria, Boente-CV performs best for the robust smoothers, except for the MMH. It is even slightly better than high breakdown point criteria like median- and 50%-LTS-CV. So the use of a median-smoother with Boente-CV is recommended for situations with an arbitrary number of outliers and without large jumps, as this combination delivers good results for moderate and large percentages of contamination. Fig. 2.4 illustrates the MASE-values for the running median and Lowess for an increasing number of outliers.

2.4.3 Robust smoothing with jumps

Next we look at situations with jumps. We take $n = 200$ observations (x_i, y_i) and include $m \in \{1, 2, 5\}$ jumps of height $s \in \{1, 3, 6\}$. For $m = 2$ the jumps are located at $\xi_1 = 0.4$ and $\xi_2 = 0.6$ and for $m = 5$ we fix $\xi_1 = 0.2, \xi_2 = 0.4, \xi_3 = 0.55, \xi_4 = 0.7$, and $\xi_5 = 0.85$. At first we only include some small outliers, fixing $\pi = 0.01$ and $\gamma = 3$.

Fig. 2.5 (left) shows that NN-median performs worse than the MW-median, due to the worse jump-preservance of a NN-method. From the jump-preserving smoothers, the MW-median and the DWMTM perform well for moderate jumps, but the un-

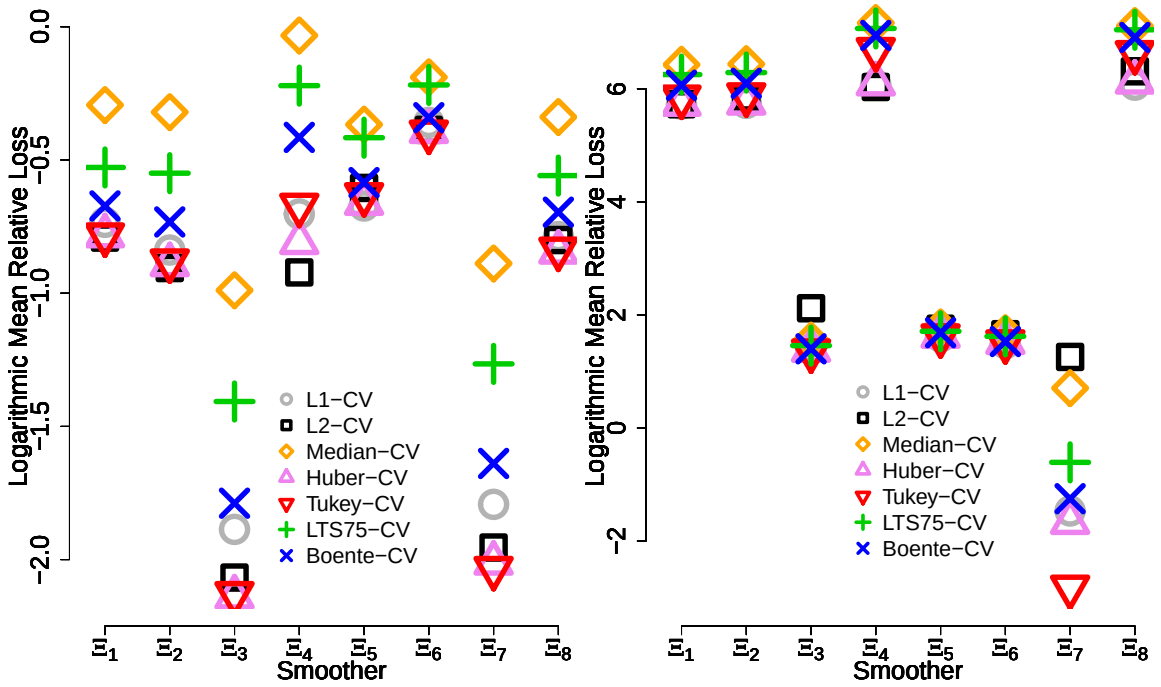

Figure 2.5: MRL for different jump situations with moderate and large jumps. DWMTM with Tukey- or Huber-CV performs best for moderate and large jumps.

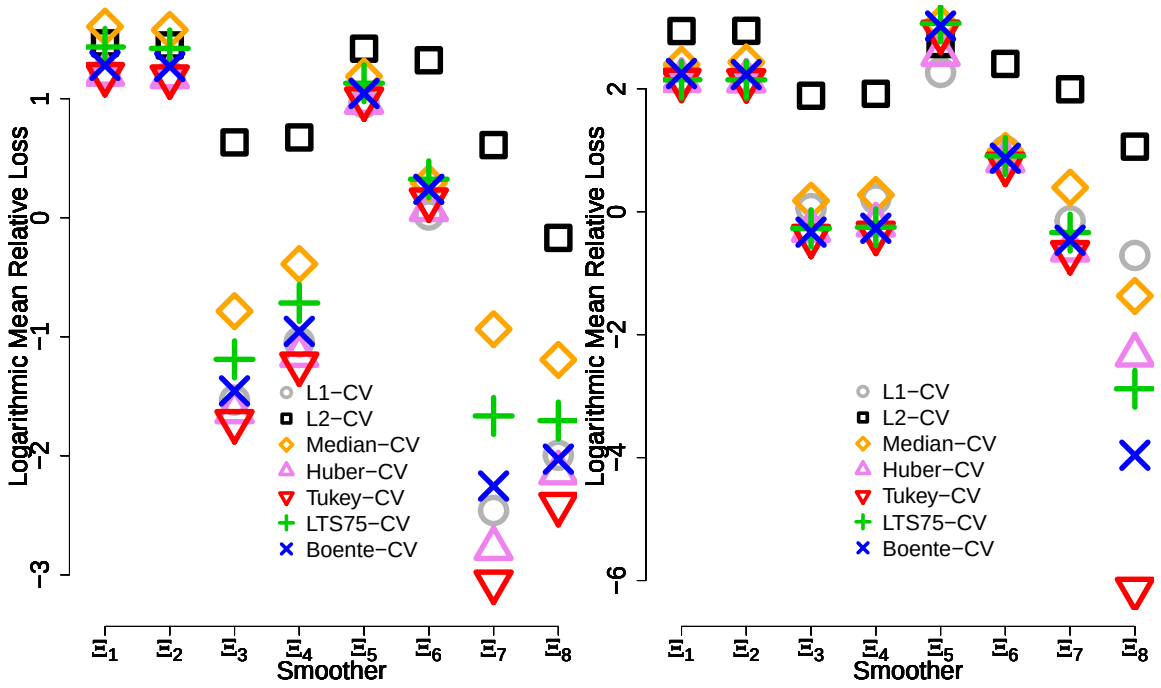

Figure 2.6: MRL for different jump situations with 5 and 15% outliers. Tukey-CV delivers good results for smoothing with jumps and outliers.

expected bad performance of the hybrid smoothers leads us to look at larger shifts. We consider $m \in \{1, 2, 5\}$ jumps of height $s \in \{12, 24, 48, 96\}$, see Fig. 2.5 (right). Although the MW-median and the two hybrid smoothers are jump-preserving, too, the DWMTM delivers a much smaller loss. While L_2-CV seems to break down for an increasing jump height, M-CV-criteria like L_1-, Huber- and Tukey-CV deliver good results for an arbitrary jump size.

Now we return to situations with smaller jumps, compare Fig. 2.5 (left), including higher percentages of larger outliers. We take $\pi = 0.05$ and $\pi = 0.15$ percent of outliers, respectively, each with a magnitude of $\gamma = 12$ so that the outliers are larger than the jumps. See Fig. 2.6 for the results. Tukey-CV is the best choice for both percentages and is preferable for situations with jumps and a moderate number of outliers. While for $\pi = 0.05$ L_1-CV also delivers good results, we observe for $\pi = 0.15$ again the superiority of a robust CV-criterion, like Boente- and LTS-CV.

The bad performance of the jump-preserving smoothers compared to Lowess in Fig. 2.6 (right) is mostly due to an incorrect jump detection because of outliers. Fig. 2.7 shows two examples of the same situation, including many large jumps and outliers. In the left example the DWMTM performs better than Lowess due to its better jump preservation, while in the right example Lowess is superior due to a wrongly tracked jump by the DWMTM. For robust jump-preserving smoothing with only a small percentage of outliers, the DWMTM with Huber- or Tukey-CV leads to good results. For larger percentages of outliers there is a trade-off between jump-preservation and outlier-robustness. If possible, one should specify lower limits for k based on the minimal number of observations in between subsequent shifts.

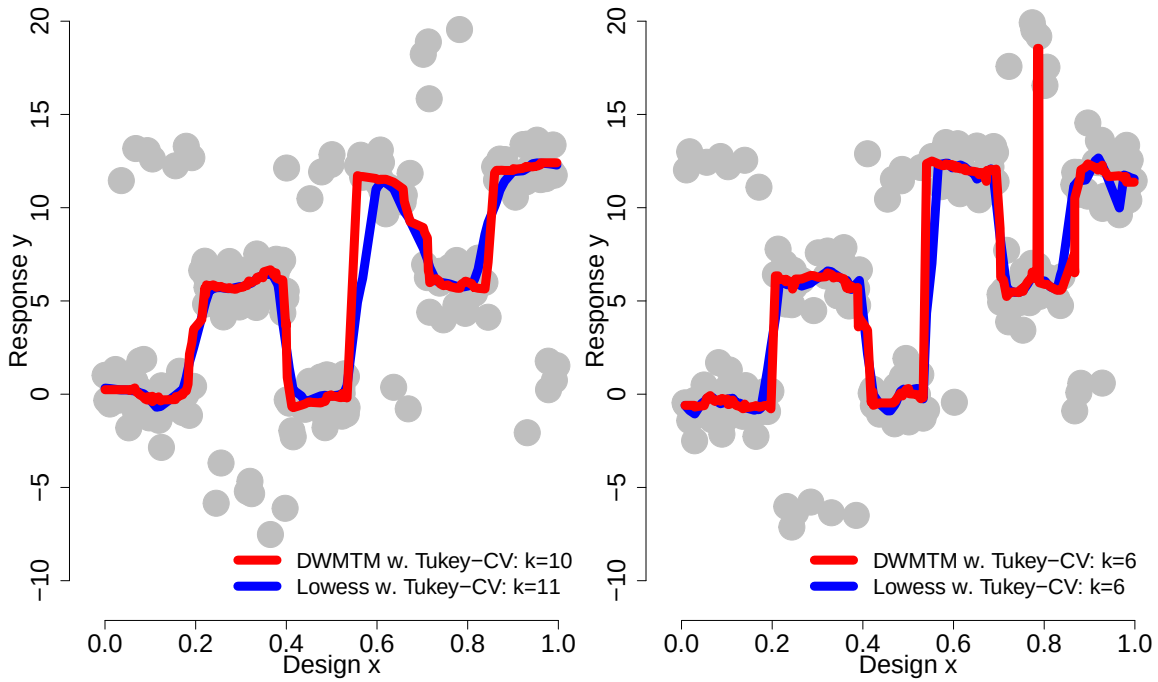

Figure 2.7: Two data examples of piecewise constant functions with 15% outliers of magnitude $\gamma = 12$ and $m = 5$ jumps with height $s = 6$.

2.4.4 Performance for different sample sizes

An open question is if the cross-validated smoothers deliver consistent estimations of the regression function f in the presence of jumps. While much is known about consistency of linear and robust smoothers if f is smooth, the asymptotic behaviour of these estimators is rather unknown if this assumption is violated. Mueller (2002) investigated the asymptotic behaviour of robust estimators, which are asymptotically linear, at jump positions and derived rather strong conditions under which robust estimators like the median or the LTS-estimator are consistent there. For these results a window width $h = n^{-1/3}$ is assumed, what is not guaranteed for a cross-validated window choice. Furthermore Hillebrand and Mueller (2006) showed that the function estimations of redescending M-kernel-smoothers can be consistent even at jump locations under some strict assumptions.

To investigate if the cross-validated smoothers are consistent we perform more simulations. We take a data set with 35 observations, 15% outliers of magnitude 192 and one jump of height 6. Then an increasing number of observations from the same model is added and the resulting sequence of estimations at the original 35 points is analysed. Note that the jump is fixed for all sample sizes, while the positions of additional outliers are randomly chosen.

Fig. 2.8 (left) shows that the performance of the MW-median combined with the L_1-CV-criterion can still be strongly influenced by outliers for $n = 200$, as for this data set short sequences of outliers are two times regarded as jumps. This leads to the short choice of $k = 2$, due to a nearly constant CV-criterion-value $LADCV(k)$.

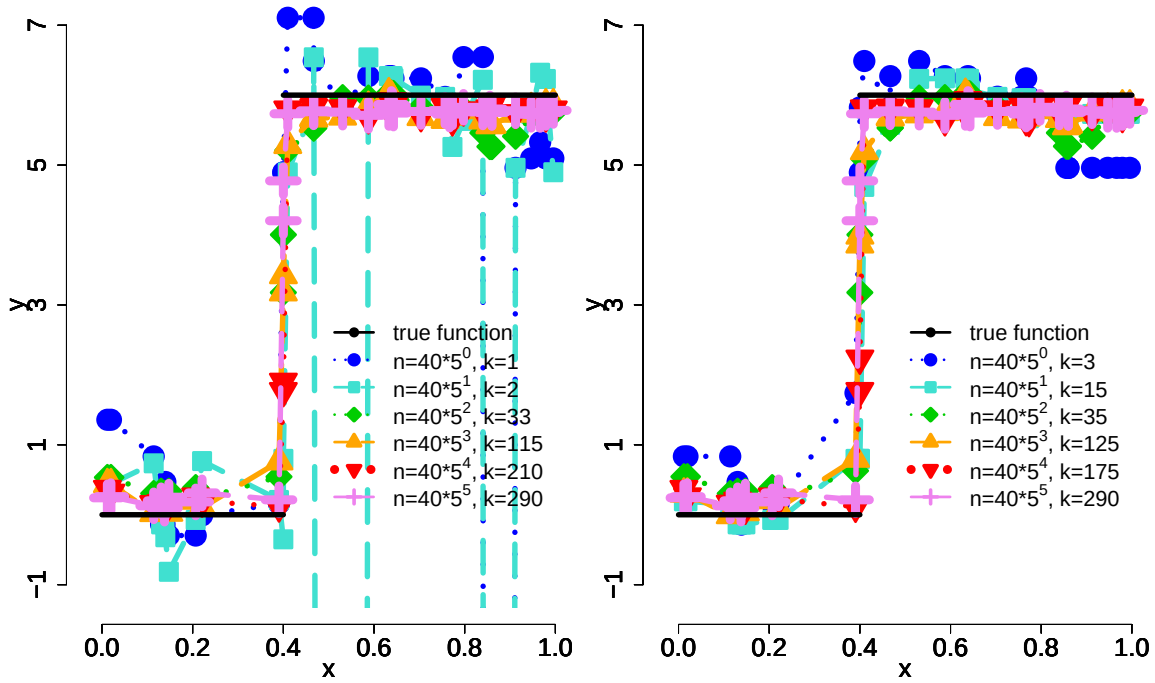

Figure 2.8: MW-median with L_1- (left) and Boente-CV (right) for an increasing n over the whole support. L_1-CV is sensitive to outliers for $n = 200$.

This phenomenon can generally be observed more often for L_1-CV than for robust CV-criteria, what is also an explanation for the worse performance of L_1-CV in the situation with large outliers (compare Fig. 2.1). Robust CV-criteria like Tukey-CV (not shown) and Boente-CV, see Fig. 2.8 (right), deliver for this data set larger values with $k = 10$ and $k = 15$, respectively. For the given data, these criteria lead to an estimation, which is not influenced by the outlying sequences. For larger n the performance of L_1-CV gets better, leading to consistent estimations at outlying points. For $n > 25000$, L_1-, Tukey- and Boente-CV deliver the same values of k.

On each side of the jump there is one design point with a distance of about 0.1 and one with a distance of nearly 0.0001 to the jump. Fig. 2.9 (right) shows that for the two points with a distance of 0.1 the MW-median achieves good results for $n \geq 5000$. This is not the case for the two design points closer to the jump, where even for $n = 125000$ the estimated function values are far away from the true ones. This indicates possible problems of consistency at the jump positions. The same is observed for the MMH and the DWMTM, independently of the CV-criterion.

To further study the behaviour for an increasing sample size n, we consider $\nu = 100$ data sets for each of $n \in \{50, 55, \ldots, 100, 110, \ldots, 200, 220, \ldots, 400\}$ and look for two data situations at the MASE-values of the different smoothers in dependence of n. See Fig. 2.10 for one situation with jumps and one with jumps and additional large outliers. The L_2-CV is used for the mean smoothers and Tukey-CV for the others. As expected all MASE-values decrease for an increasing n.

In the situation without large outliers, compare Fig. 2.10 (left), the two hybrid smoothers stay best for all n, while for $n > 300$ the DWMTM is only slightly worse,

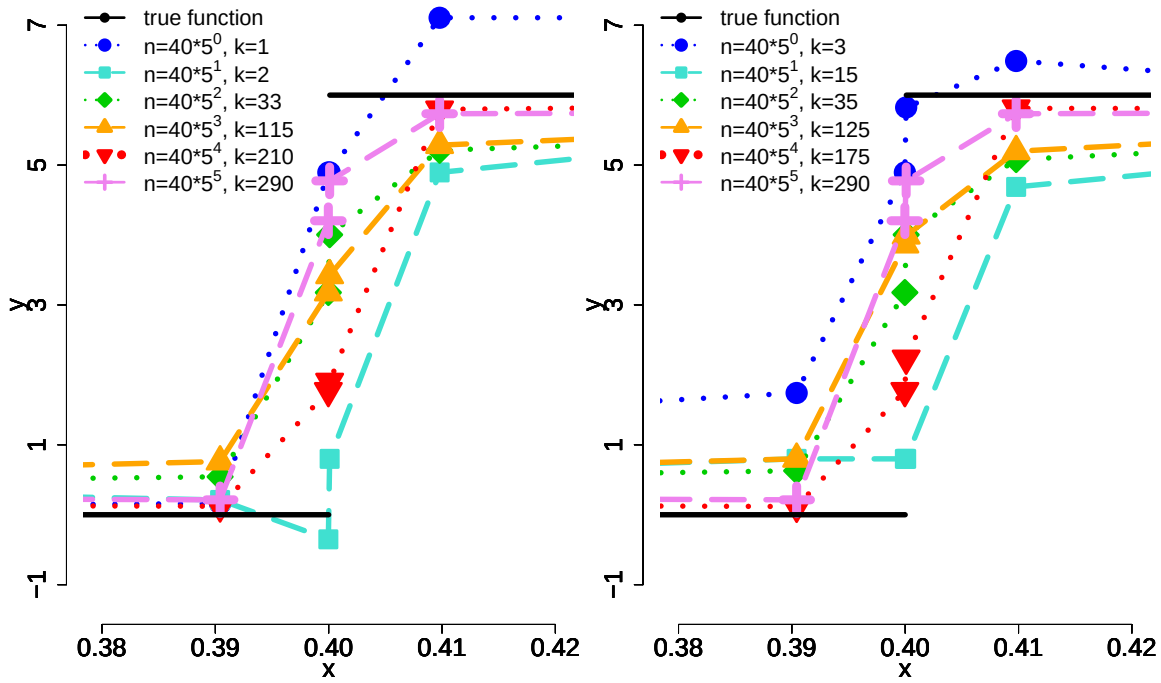

Figure 2.9: MW-median with L_1- (left) and Boente-CV (right) for an increasing n near the jump. At the jump locations, cross-validated smoothers seem to be inconsistent.

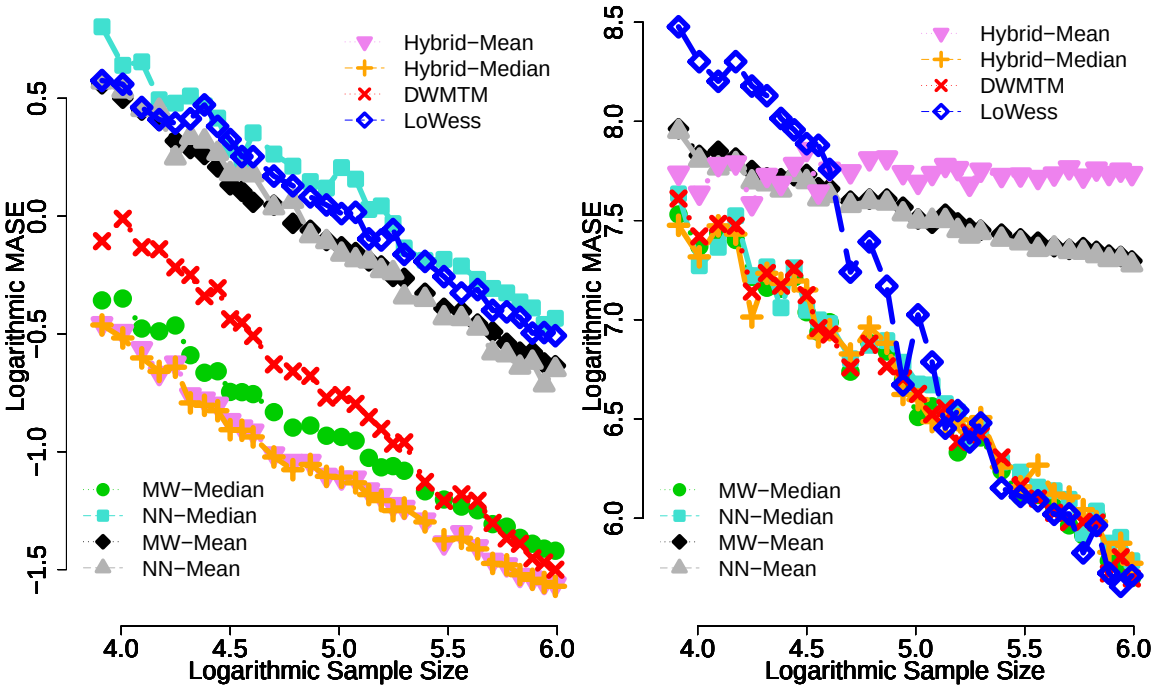

Figure 2.10: Changes in the MASE for different n for situations with 5 jumps of height 6 and without large outliers (left) or with 15% outliers of magnitude 192 (right).

indicating that especially for smaller samples the hybrid smoothers show a better

jump preservance than the DWMTM. Fig. 2.10 (right) shows for the situation with large outliers that the performance of Lowess becomes better for an increasing n, as it performs worst for small n and becomes best for $n = 400$. One explanation is that for a larger sample it is more probable that a short sequence of outliers will be falsely detected as additional jump by jump-preserving smoothers, leading to larger average squared distances between true and estimated function.

However, the MASE-criterion penalises estimators, which deliver good results for the majority of the data sets, but fail occasionally, e.g. due to incorrectly detected jumps. This becomes clear from Fig. 2.11 and 2.12, where the results for the same data situations are presented, if the Mean Averaged Absolute Error (MAAE) and the Median Median Absolute Error (MMAE), respectively, are used as performance measure. Jump-preserving smoothers sometimes detect false jumps, leading to large squared distances between true and estimated function there. In terms of the MAAE, the DWMTM becomes the best smoother for $n \geq 140$ in the situation without outliers and for $n \geq 240$ in the situation with outliers, see Fig. 2.11. In terms of the MMAE, the DWMTM is the best smoother for all sample sizes, so using the MASE as criterion has obviously disadvantaged the DWMTM, while Lowess has been favored.

The computation times needed by the several smoothers and CV-criteria are another interesting point, see Fig. 2.13. We use R (R Development Core Team, 2011) for our simulations. All functions for smoothers and CV except Lowess are self-made implementations. Faster update-algorithms could be used for smoothing with a fixed smoothing parameter, see Fried et al. (2006) for the LMH and the MMH and Bernholt

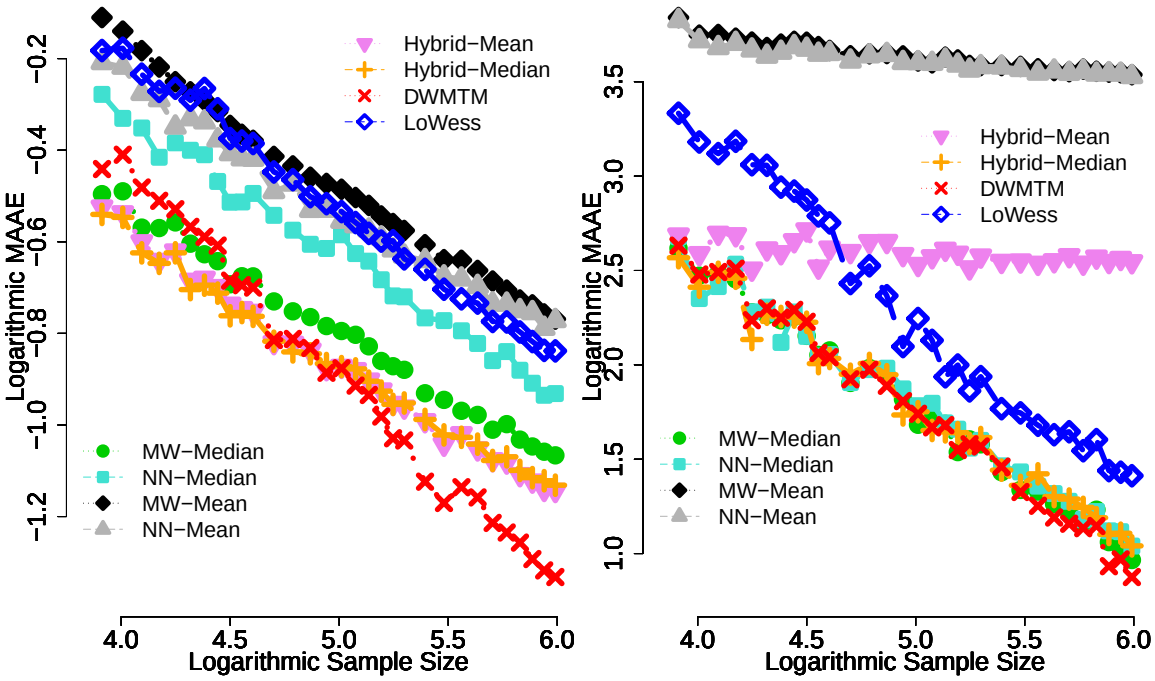

Figure 2.11: Changes in the MAAE for different n for situations with 5 jumps of height 6 and without large outliers (left) or with 15% outliers of magnitude 192 (right).

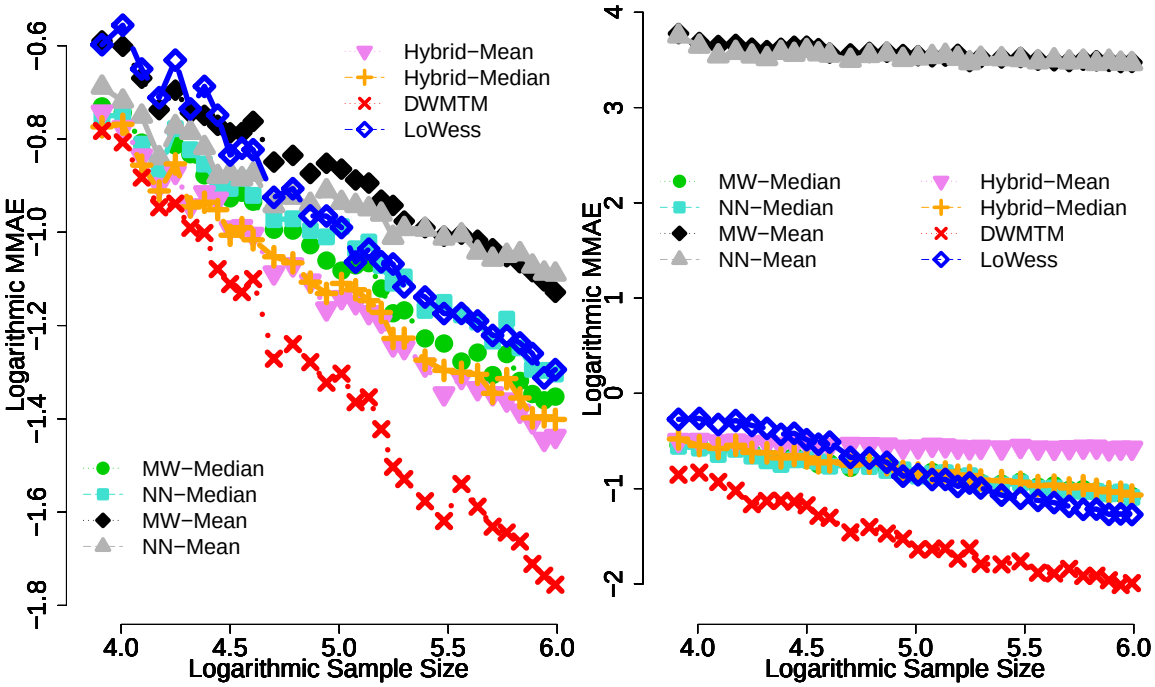

Figure 2.12: Changes in the MMAE for different n for situations with 5 jumps of height 6 and without large outliers (left) or with 15%outliers of magnitude 192 (right).

et al. (2006) for the DWMTM. Note that both axes in Fig. 2.13 are on a logarith-

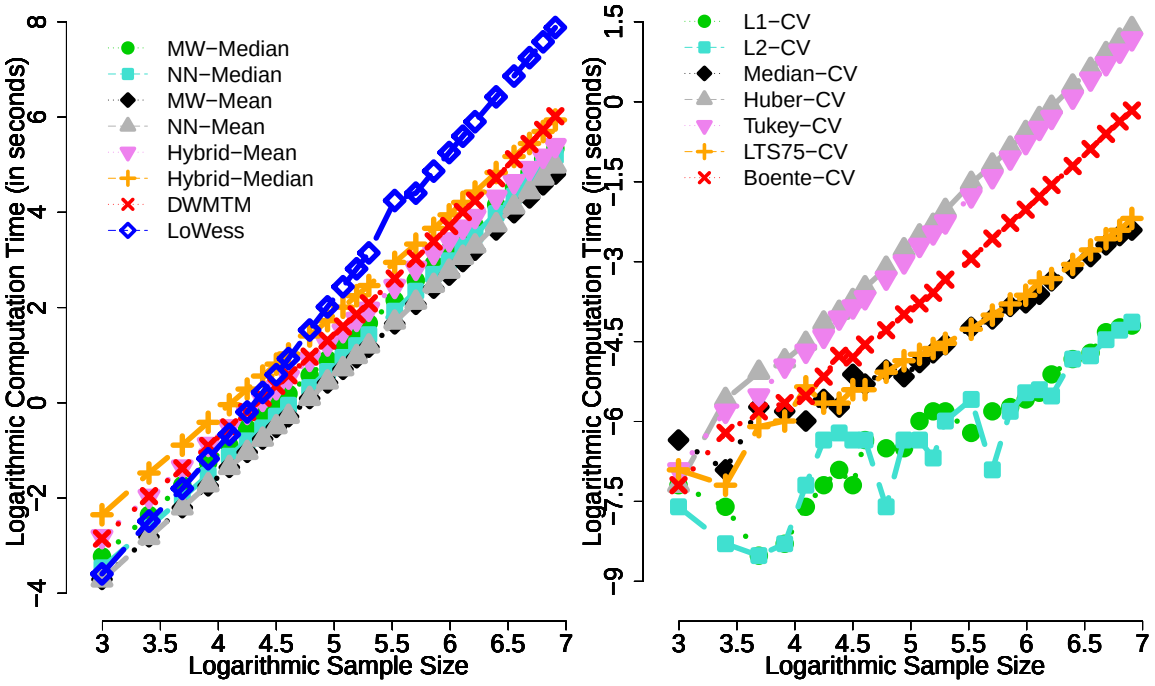

Figure 2.13: Computation times for the different smoothers and CV.

mic scale. For a large n the computation time of Lowess increases stronger than those of the other smoothers. The larger computation times for Huber-, Tukey- and Boente-CV are due to the calculation of the Q_n.

2.4.5 Performance for regression functions with curvature

Next we consider functions which are not piecewise constant. We modify the HeaviSine-function (Donoho and Johnstone, 1994), by including the amplitude α of the sine function as additional parameter, $f(x) = 4\sin(\alpha\tilde{\pi}x)$, where $\tilde{\pi} = 3.1416$.

The parameters π and γ for percentage and magnitude of the outliers as well as m and s for number and height of the jumps are regulated as in Section 2.4.2 and 2.4.3, respectively, so that we have the same simulation design here, using another class of regression functions. The situations from Section 2.4.2 and 2.4.3 can be interpreted as the special case $\alpha = 0$. The sign of the jump height is always the same as the sign of the slope of f at the jump position. We consider $\alpha \in \{1, 2, 3, 4, 5\}$. Fig. 2.14 shows two data examples, each with $\pi = 0.15$, $\gamma = 12$, $m = 5$ and $s = 6$ and amplitudes $\alpha = 2$ (left) and $\alpha = 5$ (right). Beside the jump-preservance, which is more difficult in the presence of outlying sequences, the curvature seems to be an additional problem, especially for the local constant DWMTM, what can for example be seen at the boundaries of the true function.

Lowess, as a local linear procedure, achieves the smallest losses for this new class of regression functions. For Lowess the Tukey-CV-criterion is again appropriate, just like it was in the constant function case. For the local constant smoothers an

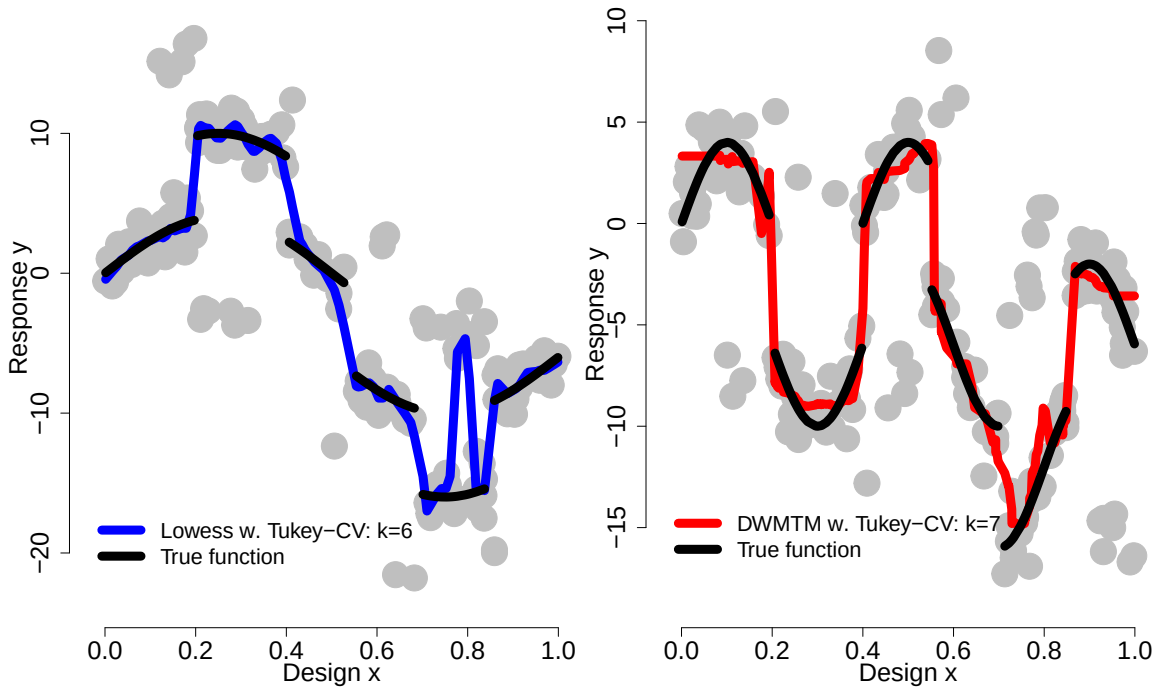

Figure 2.14: Two data examples of the modified HeaviSine-function with 15% outliers of magnitude 12, five jumps of height 6 and amplitudes $\alpha = 2$ (left) and $\alpha = 5$ (right).

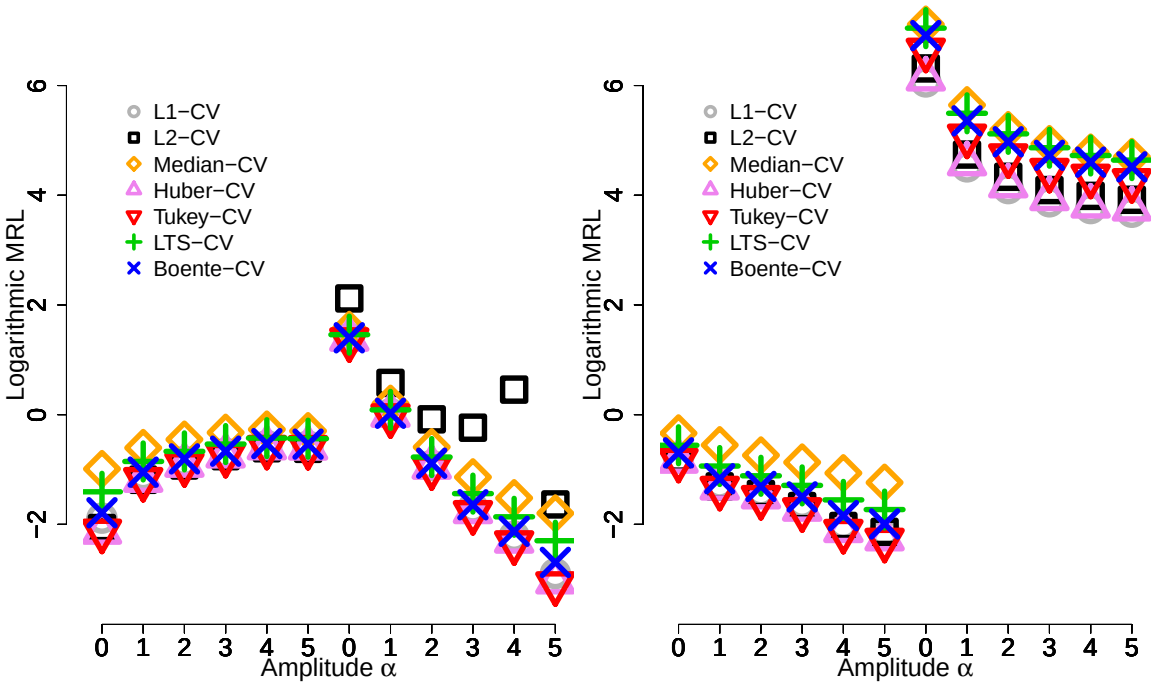

Figure 2.15: MRL-values for an increasing amplitude for moderate (first 5 amplitudes) and large (second 5 amplitudes) jumps for the MW-median (left) and Lowess (right).

unbounded CV-criterion like L_1-, L_2 or Huber-CV now delivers better results than Tukey-CV for situations, where the sine function has a large amplitude and thus a greater impact than the outliers and the jumps. Lowess is still outperformed by robust local constant smoothers like the MW-median, the MMH or the DWMTM in case of large jumps or large percentages of outliers, even if the amplitude is large. For these situations, where outliers or jumps have a greater impact than α, Tukey-CV is again appropriate for the jump-preserving smoothers. See Fig. 2.15 for the MRL-values of the MW-median and Lowess for small and large jumps and an increasing amplitude.

2.5 Real data analysis

Finally the cross-validated smoothers are applied to three real data sets. The results will be shown only for two smoothers in each case.

The well-log data (O Ruanaidh and Fitzgerald, 1996) are measurements of the nuclear magnetic response of underground rocks at 4050 time points. The underlying function f looks piecewise constant with several jumps. Each sequence belongs to a stratum of a single type of rock. There are also some short patches of large outliers. While O Ruanaidh and Fitzgerald fixed the number of jumps and removed the outliers before the data analysis, Fearnhead and Clifford (2003) used hidden Markov models to distinguish outliers from jumps automatically with the help of hyperparameters, which need prior knowledge about the magnitudes to discriminate between jumps and outliers. Robust jump-preserving smoothers also work automatically, but do not

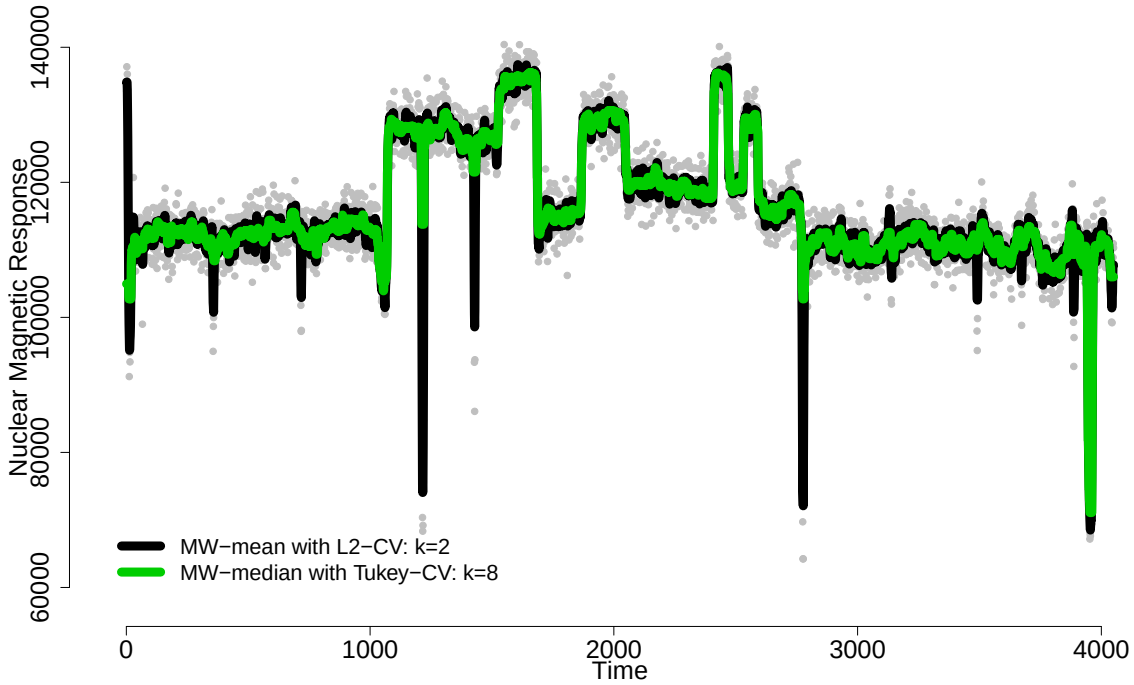

Figure 2.16: Well-log data with a robustly and a non-robustly estimated regression curve.

need such prior assumptions. Fig. 2.16 shows the data with a fit of the MW-median combined with Tukey-CV and the MW-mean with L_2-CV. As expected the median preserves all important jumps and is almost uneffected by short patches of outliers, while the MW-mean breaks down at outlying sequences. We have taken the median here, because it works well in situations with jumps and large outliers. Again, Tukey-CV shows good performance in such situations.

The second data set consists of 100 observations of the annual volume of discharge from the Nile, measured from 1871 to 1970 at Aswan in Egypt. A local constant fit seems to be adequate. Cobb (1978) recognized a decrease in the water level after 1898. There are some moderate outliers in the data. Using the results from Section 2.4.3 the MW-Median and the DWMTM, both with Tukey-CV, are jump-preserving smoothers, which can possibly handle this situation. Comparing Fig. 2.17 (left), both smoothers preserve the jump after 1898 (vertical solid line) well, but the DWMTM gives a smoother curve, especially after the jump, due to the small smoothing parameter which is proposed by the Tukey-criterion for the MW-median.

Finally we analyse the motorcycle impact data of Schmidt et al. (1981). The 133 observations are taken from a study, where the effectiveness of the helmets in collisions is determined. Time is measured in milliseconds (ms), head acceleration is measured in units of g, which is 9.8 meters per second. This is a challenging situation as the design is random, the noise heteroscedastic and the underlying function has a strong slope at about 20 ms. We show the results for the DWMTM with 75%-LTS-CV and Lowess with Tukey-CV in Fig. 2.17 (right). Both procedures give an adequate fit. Even though we expect an adaptive approach to choose local window widths instead

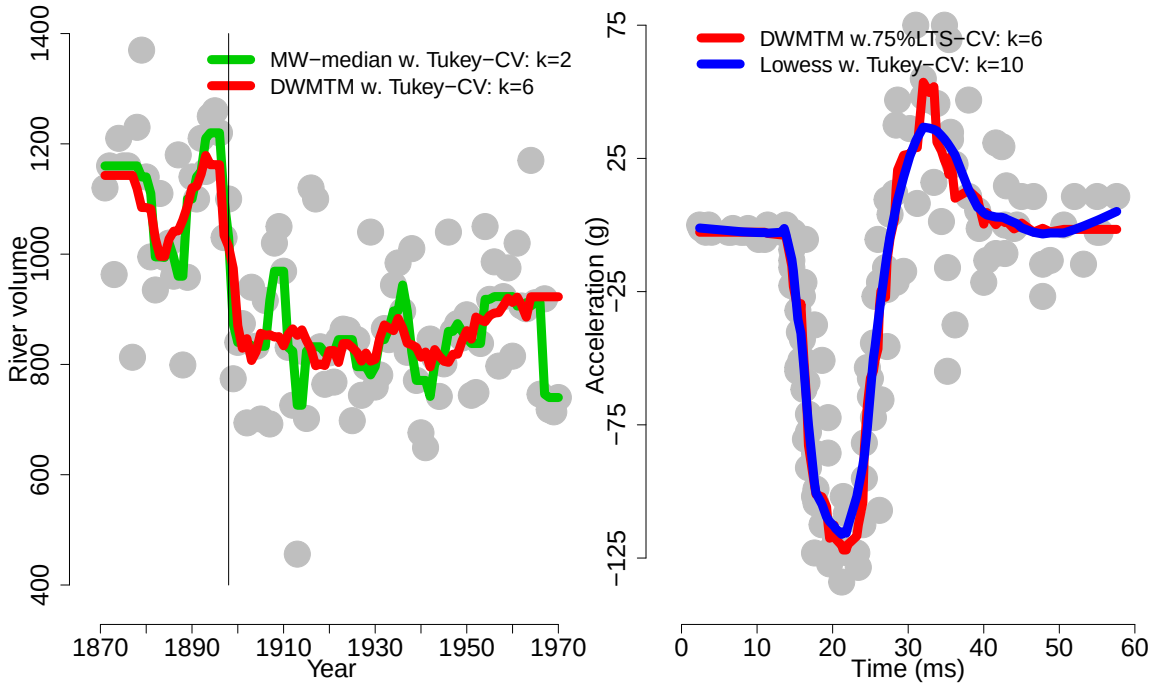

Figure 2.17: Nile data (left) and motorcycle data (right), both with estimated regression curves of the DWMTM and a robust competitor.

of a global one to perform even better for this data situation, a global width chosen by classical leave-one-out-CV based on the Tukey-criterion already delivers rather good results, except for the local extreme at about time 22. The fit of the DWMTM seems to be better here, due to the smaller window width chosen by 75%-LTS-CV.

2.6 Conclusions

Jumps and outliers are challenges for smoothers. Both, smoothing method and cross-validation (CV) criterion need to be chosen properly then. L_1-CV is often considered to be a robust CV-criterion, but like L_2-CV, it can lead to poor results in the presence of large outliers. In the absence of large jumps, Boente-CV with the Q_n scale estimator leads to the best results among the robust criteria, as it deals well even with a large number of outliers. If the regression function includes discontinuities, CV-criteria based on Huber's or Tukey's loss function work well for jumps of arbitrary size. Tukey-CV even delivers good results if the data are additionally contaminated by outliers. Use of L_2-CV is discouraged in the presence of large outliers or jumps.

From the smoothers considered here, Lowess with a robust CV-criterion is to be recommended, if the contamination by outliers is moderate and the underlying constant function has no relevant jump discontinuity. For more than 25% outliers or for large jumps, Lowess loses its superiority and a jump-preserving smoother like the double window modified trimmed mean (DWMTM) is preferable. This even applies to curved functions like the sine.

Multivariate extensions are possible, but computationally expensive, as one should allow different window widths for all regressors. Not only the CV-criteria, but also a reasonable preselection of tuples of bandwidths has to be considered then. Alternatively, existing approaches for window width selection for multivariate regression (Kerkyacharian et al., 2001; Lafferty and Wasserman, 2008; Zhang et al., 2009) or regression with functional data (Benhenni et al., 2007) could be robustified.

Further investigations could be made considering the window width selection for local linear smoothers in the presence of outliers and jumps. Although our procedures worked well for the motorcycle impact data, local linear procedures should work better for functions with larger slopes. There are many robust regression methods (Davies et al., 2004; Gather et al., 2006), which could be combined with the robust CV-criteria considered here. If additionally the error terms are dependent, like in time series, there are alternative types of CV to the classical leave-one-out-version (Francisco-Fernandez and Vilar-Fernandez, 2005), which can also be robustified and compared. This can provide improvements e.g. for robust time series analysis.

Chapter 3

On robust nonparametric smoothing: The indirect approach

3.1 Introduction

We consider a regression model

$$Y_i = f(X_i) + E_i,\ i = 1, \ldots, n\ , \tag{3.1}$$

where f is an unknown piecewise continuous function, $x_1, \ldots, x_n$ are values of a covariate generated from a random design $X_1, \ldots, X_n$,$E_1, \ldots, E_n$ are i.i.d. normal distributed errors, possibly contaminated by some outliers, and $Y_1, \ldots, Y_n$ are observations of a response variable measured at $x_1, \ldots, x_n$, with realisation $y_1, \ldots, y_n$. For simplicity of the exposition we assume the data to be ordered according to the size of the x_i, $x_1 \leq x_2 \leq \ldots \leq x_n$.

In this Chapter we want to find good estimations of the jump locations $\xi_1, \ldots, \xi_m$ of the true function f. These estimations should accomplish the following goals:

(i) The estimated number of jumps $\widehat{m}$ should coincide with the true m.
(ii) The estimated jump locations $\widehat{\xi}_1, \ldots, \widehat{\xi}_{\widehat{m}}$ should be close to $\xi_1, \ldots, \xi_m$.
(iii) The resulting estimation $\widehat{f}$ should give an appropriate fit of f.

These three goals should not be treated separately, because they are interdependent. Especially (i) and (ii) should be researched jointly, as a jump detection rule is only good, if the true number of jumps is estimated correctly and if the true jump locations are determined at the same time.

The proposed jump detection rules in this Chapter are constructed in the same way, which Gijbels et al. (2007) describes as indirect approach: In a first step the jump locations $\xi_1, \ldots, \xi_m$ are estimated, what also delivers an estimation of the number of jumps m. In a second step the whole support is splitted into several parts, which are separated by the estimated jump locations. In each part of the support, f is assumed to be continuous and estimated by a global location estimator.

To find good selection rules for different situations with outliers, we use an approach based on test statistics, which compare the location parameters between two

samples. To test, if the function f has a jump discontinuity between the design points x_i and x_{i+1}, $i = k, \ldots, n-k$, we assume that the observations $y_{i-k+1}, \ldots, y_i$ are a realisation of i.i.d. random variables with location parameter μ_-^i and analogously $y_{i+1}, \ldots, y_{i+k}$ are a realisation of i.i.d. random variables with location parameter μ_+^i. Then rejection of H_0^i in the test problem

$$\mathrm{H}_0^i: \ \mu_-^i = \mu_+^i \ \text{ vs. } \ \mathrm{H}_1^i: \ \mu_-^i \neq \mu_+^i \ \ , i \in \{k, \ldots, n-k\} \text{ fixed} \tag{3.2}$$

indicates a jump discontinuity. This can be repeated for each design point x_i, except $x_1, \ldots, x_{k-1}, x_{n-k+1}, \ldots, x_n$, because we have not enough observations to the left of x_{k-1} and to the right of x_{n-k+1}, respectively. Which location measure and thus which test statistic to use depends on further assumptions on the data. Furthermore for a fixed n we have a different number of test problems (3.2) for different choices of k. Thus a multiple adjustment of the level of significance by the method of Bonferroni or Bonferroni-Holm can possibly bring an improvement over a fixed chosen level for each test, which is independent of k.

The subscription of the observations shows that the performance of the jump detection rule will depend on an appropriate choice of the (smoothing) parameter k. A too large value of k can lead to non-detection of some discontinuities, e.g. if less than k observations between two jumps are observed. Also too small choices of k can lead to the detection of too many jumps. This implies shorter sequences of observations and so more imprecise estimations of the true function in these sequences.

A proper choice of the parameter k can be made by cross-validation (CV). Morell et al. (2012) compare several robust CV-criteria in the context of jump-preserving smoothing in situations with jumps and outliers. A similar comparative study is of interest for the indirect approach described above. Till now, there is only one proposal, which uses CV for choosing k in the indirect approach. Gijbels and Goderniaux (2004a,b) take the derivative of the Nadaraya-Watson-estimator as a diagnostic function to obtain a first rough estimation of the jump discontinuity. This rough estimation is then improved by a smooth estimation based on least squares in a shorter interval around the estimated jump location. L_2-CV is used to achieve the window width with the maximal absolute value of the derivative in the first step and the length of the interval in the second step. However, in the presence of outliers, L_2-CV delivers an inappropriate choice of k and robust CV-criteria are needed. The behaviour of those robust CV-criteria has obviously not been investigated for the indirect approach, yet.

Another problem in this context is that one jump can induce the rejection of more than one null hypothesis H_0^i. Assume that we have a jump discontinuity between the design points x_i and x_{i+1}. The null hypothesis H_0^j may not only be rejected for $j = i$, but also for other values of j in the neighborhood of i. Therefore Wu and Chu (1993) and Qiu and Yandell (1998) proposed rules to detect the most likely jump location points out of all candidates under the assumption that candidates, which are too close to each other, belong to the same jump discontinuity, see Section 3.2.3 for details. We also consider slightly modifications of the two original rules, which bring advantages in most data situations.

Altogether there are several questions belonging to the indirect approach:

(a) Which two sample test statistics and corresponding smoothers work well in robust jump detection, when outliers are observed?
(b) What is preferable for the explorative test procedures: A multiple level of significance for the whole data set or a fixed level at each design point?
(c) Which procedure for finding the most likely candidates for a jump should be used: The approach of Wu and Chu or the one of Qiu und Yandell?
(d) Which cross-validation criteria are appropriate to improve the jump detection?
(e) Which criteria provide information about the quality of a jump detection rule?

We will try to answer these questions in the following. Section 3.2 gives informations about the single detection rules and Section 3.3 will show the results of several simulation studies. Section 3.4 finally concludes.

3.2 Description of the used detection rules

3.2.1 Detection rules: Test statistics

In a first step we introduce the explorative use of the two sample test statistics. For a fixed $i \in \{k, \ldots, n-k\}$ we define two samples

$$\begin{aligned} \mathcal{S}_-^i &= \{Y_{i-k+1}, \ldots, Y_i\} =: \{Y_{1;-}^i, \ldots, Y_{k;-}^i\} \text{ and} \\ \mathcal{S}_+^i &= \{Y_{i+1}, \ldots, Y_{i+k}\} =: \{Y_{1;+}^i, \ldots, Y_{k;+}^i\}, \end{aligned} \tag{3.3}$$

which consist of k i.i.d. random variables (rv), respectively. Further the rvs $Y_{i-k+1}, \ldots, Y_{i+k}$ are independent altogether. Regarding model (3.1) and assuming additionally that all expectations exist, we have

$$\mathrm{E}(Y_j) = \underbrace{\mathrm{E}(f(X_j))}_{=\mu_-^i} + \underbrace{\mathrm{E}(E_j)}_{=0} = \mu_-^i \text{ and } \mathrm{E}(Y_{j'}) = \underbrace{\mathrm{E}(f(X_{j'}))}_{=\mu_+^i} + \underbrace{\mathrm{E}(E_{j'})}_{=0} = \mu_+^i \ , \tag{3.4}$$

respectively, for $j = i-k+1, \ldots, i$ and $j' = i+1, \ldots, i+k$. We check the null hypothesis H_0^i of the test problem

$$\mathrm{H}_0^i : \ \mu_-^i = \mu_+^i \ \text{ vs. } \ \mathrm{H}_1^i : \ \mu_-^i \neq \mu_+^i \tag{3.5}$$

by testing whether a level shift occurs between the design points X_i and X_{i+1}. A popular choice for this problem is the ordinary two sample t-test

$$T_1 = \frac{|\overline{Y}_-^i - \overline{Y}_+^i|}{\widehat{S}^i \sqrt{\frac{2}{k}}}, \tag{3.6}$$

$$\text{with} \quad \overline{Y}_-^i = \frac{1}{k} \sum_{j=1}^{k} Y_{j;-}^i \quad \text{and} \quad \overline{Y}_+^i = \frac{1}{k} \sum_{j'=1}^{k} Y_{j';+}^i \tag{3.7}$$

being the two sample means and the pooled empirical variance

$$\widehat{S^i}^2 = \frac{1}{2(k-1)} \left(\sum_{j=1}^{k} (Y_{j;-}^i - \overline{Y}_-^i)^2 + \sum_{j'=1}^{k} (Y_{j';+}^i - \overline{Y}_+^i)^2 \right) \tag{3.8}$$

within the two samples. Under Gaussian noise, T_1 follows a t-distribution with $2(k-1)$ degrees of freedom and the threshold value t_1 is the $(1-\frac{\alpha}{2})$-quantile of this distribution, with $\alpha \in (0,1)$ being the level of significance. In this situation the t-test is the best unbiased test for H_0^i vs H_1^i.

If outliers occur, the t-test can easily lead to a wrong decision. Ylvisaker (1977) defines the concept of resistance to acceptance and rejection for a given test as a similar concept of robustness like the breakdown point ε_N for estimators, $N = 2k$. The finite sample breakdown point corresponds to the minimal fraction of modifications in a sample which can drive the estimate to the boundaries of the parameter space.

Similarly, the resistance to acceptance ε_0 corresponds to the minimal fraction k_0 of modifications in the entire sample $Y_{i-k+1}, \ldots, Y_{i+k}$, which leads to the acceptance of H_0^i, independently of the $2k - k_0$ other observations. For T_1 we have $\varepsilon_0(T_1) = \frac{1}{2k}$, since for every k and α one outlier can change a mean to any value and can lead to a numerator of zero and to acceptance, independently of the other observations.

Analogously the resistance to rejection ε_1 is defined as the minimal fraction k_1 of modifications in the entire sample $Y_{i-k+1}, \ldots, Y_{i+k}$, which leads to the rejection of H_0^i, independently of the $2k - k_1$ other observations. This resistance is harder to determine, as large outliers will not only lead to a large difference of the means, but also to a large sample variance. Furthermore finding k_1 so that H_0^i can be rejected for all α in arbitrary samples, needs the modification of at least $k_1 = k$ observations (Fried, 2007). Nevertheless, for a given level α fewer modifications can be enough to exceed the critical value, see Ylvisaker (1977) for $\alpha = 0.05$, where $k_1 = 4$ outliers are already enough to reject H_0^i for all $k \geq 4$.

Several robust alternatives to the t-test are presented in the literature. One idea is to replace the low breakdown-point estimators $\overline{Y}^i_-$, $\overline{Y}^i_+$ and $\widehat{S}^i$, respectively, by robust estimators. Yuen and Dixon (1973) introduced a trimmed t-test:

$$T_2 = \frac{|\overline{Y}^i_{\varpi;-} - \overline{Y}^i_{\varpi;+}|}{\widehat{S^i_\varpi}\sqrt{\frac{1}{(k-2\vartheta)(k-2\vartheta-1)}}}, \tag{3.9}$$

with $\varpi \in (0, 0.5)$ being a trimming factor, defining the number $\vartheta = \lfloor k\varpi \rfloor$ of observations, which is trimmed away, and correspondingly the trimmed means

$$\overline{Y}^i_{\varpi;-} = \frac{1}{k-2\vartheta}\sum_{j=\vartheta+1}^{k-\vartheta} Y^i_{(j);-} \text{ and } \overline{Y}^i_{\varpi;+} = \frac{1}{k-2\vartheta}\sum_{j'=\vartheta+1}^{k-\vartheta} Y^i_{(j');+} \tag{3.10}$$

as well as the winsorised sum of squared deviations to the trimmed means

$$\begin{aligned}\widehat{S^i_\varpi}^2 &= \vartheta(Y^i_{(\vartheta+1);-} - \overline{Y}^i_{\varpi;-})^2 + \sum_{j=\vartheta+1}^{k-\vartheta}(Y^i_{(j);-} - \overline{Y}^i_{\varpi;-})^2 + \vartheta(Y^i_{(k-\vartheta);-} - \overline{Y}^i_{\varpi;-})^2 \\ &+ \vartheta(Y^i_{(\vartheta+1);+} - \overline{Y}^i_{\varpi;+})^2 + \sum_{j'=\vartheta+1}^{k-\vartheta}(Y^i_{(j');+} - \overline{Y}^i_{\varpi;+})^2 + \vartheta(Y^i_{(k-\vartheta);+} - \overline{Y}^i_{\varpi;+})^2 \ .\end{aligned}$$

The winsorised mean delivers better results than the trimmed mean at the Gaussian distribution, but for heavy-tailed distributions, where robust methods are necessary,

the trimmed mean is often more efficient (Tukey and McLaughin, 1963). For standardisation of the trimmed mean the winsorised variance is more appropriate than the trimmed variance (Tukey and McLaughin, 1963), what explains the combination from above. We choose $\varpi = 0.25$ to compromise between robustness and efficiency.

Yuen and Dixon (1973) showed that under normality for sample sizes $k \geq 7$, the empirical distribution of T_2 can be approximated by a t-distribution with $2(k-2\vartheta-1)$ degrees of freedom. So for sample sizes large enough, the realised value t_2 can be compared to the $(1-\frac{\alpha}{2})$-quantile of this t-distribution. For sample sizes $k < 7$ one should simulate the empirical $(1-\frac{\alpha}{2})$-quantiles of T_2 and compare it to the observed value of t_2. Under Gaussian noise the power efficiency of T_2 compared to T_1 is nearly $100(1-\frac{2\vartheta}{3k})\%$. Under symmetric heavy tailed noise the trimmed version delivers better results than the original t-test. The approximation seems to be appropriate for an adequate relation between trimmed sample size and proportion of outliers.

Fried and Dehling (2011) introduced three alternative tests, based on medians and Hodges-Lehmann-estimators. All tests are distribution free. The first test statistic

$$T_3 = \frac{|\widetilde{Y}^i_{0.5;-} - \widetilde{Y}^i_{0.5;+}|}{\text{Med}\left(\left\{|Y_{i-k+j} - \widetilde{Y}^i_{0.5;-}|\right\}_{\{j=1,\ldots,k\}}, \left\{|Y_{i+j'} - \widetilde{Y}^i_{0.5;+}|\right\}_{\{j'=1,\ldots,k\}}\right)} \tag{3.11}$$

is based on a comparison of medians. It is asymptotically normally distributed, but for sample sizes $k \leq 20$ Fried and Dehling (2011) suggest to use permutation tests to derive critical values for T_3. The possible number of all permutations if the sample is splitted into two parts of size k is $\binom{2k}{k}/2$ and so increases with an exponential growth rate for an increasing k. It is recommended to draw 1000 random samples, to calculate the value of T_3 for each sample and to use the $(1-\frac{\alpha}{2})$-quantile of the empirical distribution as critical value.

Fried and Dehling (2011) also considered two tests based on Hodges-Lehmann estimators (HLE; see Hodges and Lehmann 1963) instead of medians. The first is based on the difference of the two one-sample-HLEs, which are defined as

$$\begin{aligned} \widehat{Y}^i_{1;-} &= \text{Med}\left(\left\{\frac{Y_{j_1} + Y_{j_2}}{2}\right\}_{\{i-k+1 \leq j_1 < j_2 \leq i\}}\right) \text{ and} \\ \widehat{Y}^i_{1;+} &= \text{Med}\left(\left\{\frac{Y_{j'_1} + Y_{j'_2}}{2}\right\}_{\{i+1 \leq j'_1 < j'_2 \leq i+k\}}\right). \end{aligned} \tag{3.12}$$

The test statistic is given by

$$T_4 = \frac{|\widehat{Y}^i_{1;-} - \widehat{Y}^i_{1;+}|}{\text{Med}\left(\left\{|Y_{i-k+\tilde{j}} - \widetilde{Y}^i_{0.5;\tilde{j}\pm} - Y_{i-k+\tilde{j}'} + \widetilde{Y}^i_{0.5;\tilde{j}'\pm}|\right\}_{\{1 \leq \tilde{j} < \tilde{j}' \leq N\}}\right)}. \tag{3.13}$$

Note that the scale estimator in the denominator compares all pairs of the N observations of both samples, corrected with the respective sample median

$$\widetilde{Y}^i_{0.5;\tilde{j}\pm} = \begin{cases} \widetilde{Y}^i_{0.5;-}, & \tilde{j} \leq k \\ \widetilde{Y}^i_{0.5;+}, & \tilde{j} > k \end{cases} \quad \text{and } \widetilde{Y}^i_{0.5;\tilde{j}'\pm} = \begin{cases} \widetilde{Y}^i_{0.5;-}, & \tilde{j}' \leq k \\ \widetilde{Y}^i_{0.5;+}, & \tilde{j}' > k \end{cases} \tag{3.14}$$

to yield an appropriate estimation of the variability within the samples.

The next test uses the two-sample-HLE (Hodges and Lehmann, 1963):

$$T_5 = \frac{|\text{Med}(\{Y_{i-j+1} - Y_{i+j'}\}_{\{j=1,\ldots,k,\ j'=1,\ldots,k\}})|}{\text{Med}\left(\left\{|Y_{i-k+\tilde{j}} - \widetilde{Y}^i_{0.5;\tilde{j}\pm} - Y_{i-k+\tilde{j}'} + \widetilde{Y}^i_{0.5;\tilde{j}'\pm}|\right\}_{\{1 \le \tilde{j} < \tilde{j}' \le N\}}\right)}. \tag{3.15}$$

The two-sample-HLE is the median of all differences between two observations of different samples. It corresponds to the value, which is needed to obtain the same or at least a most similar rank-sum in both samples. For both tests T_4 and T_5 the permutation principle should be used again for sample sizes $k \le 12$ to derive critical values as the asymptotic test versions do not achieve approximately correct levels of significance there (Fried and Dehling, 2011).

In the following we will slightly differ from Fried and Dehling (2011) and use the two sample-MAD-scale-measure from equation (3.11) for standardisation of T_4 and T_5. It achieves better results, when large magnitudes or percentages of outliers are observed and it takes less computation time. Also, as the computation of the cross-validated smoothers would take too much time, we will not use the permutation principle, but compute exact critical values of T_3, T_4 and T_5, respectively, under the normality assumption, as in a first step we will only consider normally distributed error terms, which are possibly contaminated by outliers.

Another test based on ranks is the well-known Wilcoxon rank-sum test

$$T_6 = \sum_{j=1}^{k} \sum_{\tilde{j}=1}^{N} \mathbb{1}_{(0,\infty)}(Y_{i-k+j} - Y_{i-k+\tilde{j}}),$$

which sums the ranks of the first sample, when the ranks are calculated over all $N = 2k$ observations $Y_{i-k+1}, \ldots, Y_{i+k}$. The critical value of T_6 can be derived from the exact distribution, which is easy to calculate, because under H_0^i all possible permutations of the N ranks have the same probability.

Another linear rank statistic is the one used in the median-test

$$T_7 = \sum_{j=1}^{k} \mathbb{1}_{(0,\infty)}(Y_{i-k+j} - \widetilde{Y}^i_{0.5}), \tag{3.16}$$

which calculates the median $\widetilde{Y}^i_{0.5}$ over all N observations $Y_{i-k+1}, \ldots, Y_{i+k}$ and counts the number of observations from the first sample, which exceed $\widetilde{Y}^i_{0.5}$. T_7 follows a hypergeometric distribution under H_0^i with k drawings, successes and failures, respectively. Its $(1 - \frac{\alpha}{2})$-quantile can be used as threshold-value.

Fried and Dehling (2011) found the two tests T_4 and T_5 based on HLEs to perform better than the related linear Wilcoxon rank-statistic T_6, when the error distribution is heavy-tailed or outliers occur, while being similarly good for Gaussian noise. T_5 performed slightly better than T_4 in their simulations. Similar results were observed for the median-based tests T_3 and T_7, where the linear rank-statistic performed inferior to the median-comparison for all considered distributions.

3.2.2 Detection rules: Adjustment of the level of significance

Beside the test-statistic, there are other important questions to answer in data-based jump detection. The first important question is the value of the significance level α, which has to be chosen before the data analysis. For a sample size of n and a smoothing parameter k we have $n - 2k + 1$ test problems for the same data set and so the need of adjusting α due to a multiple test problem. Even if we use the test procedures from Section 3.2 only as an explorative tool to find candidates for the jump positions, an adjusted level chosen by the procedures of e.g. Bonferroni or Bonferroni-Holm (Holm, 1979) could lead to a better comparability of data situations with different sample sizes n or parameter values k. We compare a fixed level of significance of $\alpha_f = 0.01$ with an adjusted level α_b (by the Bonferroni-method) and α_h (by the Bonferroni-Holm-method), both with a global level of significance of $\alpha_g = 0.2$.

Given a multiple comparison with $n - 2k + 1$ test problems, the method of Bonferroni simply splits the global level α_g uniformly to all test problems, leading to $\alpha_b^i = \frac{\alpha_g}{n-2k+1}, i = k, \ldots, n-k$. So we will reject H_0^i for each i, where the p-value to the realised test statistic is smaller than α_b^i. This correction rule is conservative, leading to a smaller power of the related test and so possibly to a smaller detection rate. We will also consider the less conservative adjustment rule of Bonferroni-Holm. Here in a first step the p-values p_i of all $n - 2k + 1$ test problems are sorted and the ordered p-value $p_{(i')}$ is compared with $\alpha_h^{i'} = \frac{\alpha_g}{n-2k+2-i'}$ for all $i' \in \{1 \ldots, n - 2k + 1\}$, until for the first i' the related test can not be rejected. Then all the following tests with larger p-values will not be rejected, too.

So for H_0^i with the smallest p_i, α_b^i and α_h^i are the same, while for each of the other test problems, α_h^i will have a greater value, so that the Bonferroni-Holm-method can find more jump location candidates than the Bonferroni-method in a given data set.

3.2.3 Detection rules: Selection of the jump positions

If a jump between x_i and x_{i+1} exists, we will not only reject the corresponding Null H_0^i with high probability, but also $H_0^j, j = i - k + 1, \ldots, i - 1, i + 1, \ldots, i + k - 1$, as one of the related samples $\mathcal{S}_-^j$ and $\mathcal{S}_+^j$ includes at least one observation from the sample $\mathcal{S}_-^i$ before the jump and at least one observation from the sample $\mathcal{S}_-^i$ after the jump. Therefore we have to eliminate some detected jump-candidates. Intuitively, the probability of a rejection will be generally decreasing, if j drifts further away from i. We denote $\Upsilon_i = \{x_{i-k+1}, \ldots, x_i, x_{i+1}, \ldots, x_{i+k}\}$ as the neighbourhood of x_i and assume x_i to be the position of the true jump, if H_0^i is rejected.

Wu and Chu (1993) propose a rule to eliminate close-by jump-candidates for jump detection with kernel-estimators, which we will modify here for an arbitrary testing procedure T_a, e.g. $T_1, \ldots, T_7$, with its critical value $c_{1-\frac{\alpha^i}{2};a}$ and α^i chosen fixed or multiple adjusted by Bonferroni or Bonferroni-Holm. Let

$$\Psi_1^a = \{i \in \{k, \ldots, n-k\} : \ |T_a(\Upsilon_i)| \ > c_{1-\frac{\alpha^i}{2};a}\} \tag{3.17}$$

be the set of indices belonging to the design points, for which H_0^i can be rejected. So Ψ_1^a includes the indices of all possible candidates for a jump of f, detected via

test-procedure T_a. Following Wu and Chu (1993) we use the algorithm below for estimating the true jumps within the interval $[x_k, x_{n-k}]$, starting with $b = 1$ and assuming that $\Psi_1^a \neq \varnothing$, otherwise the algorithm will not detect any jump for the given data set. The algorithm uses the following steps:

Step 1: Find the index $\psi_b = \arg\max_{i \in \Psi_b^a} |T_a(\Upsilon_i)|$, for which the corresponding H_0^i is most clearly rejected (i.e. the index with the smallest p-value).

Step 2: Estimate the corresponding jump location $\widehat{\xi}_b = x_{\psi_b}$.

Step 3: Define a new set Ψ_{b+1}^a by $\Psi_{b+1}^a = \Psi_b^a \setminus \{\psi_b - 2k, \ldots, \psi_b + 2k\}$

Step 4: Stop if $\Psi_{b+1}^a = \varnothing$, otherwise set $b = b + 1$ and go back to Step 1.

This algorithm leads directly to estimations of the number of jumps $\widehat{m} = b$ and of the $\widehat{m}$ jump locations $\widehat{\xi}_1, \ldots, \widehat{\xi}_{\widehat{m}}$. Assuming again that a jump is detected between x_i and x_{i+1}, the subtracted value $2k$ from Step 3 causes that no other jump location will be detected in the interval $[x_{i-2k}, x_{i+2k}]$ and possible rejections of

$$\mathrm{H}_0^j, j = i - 2k, \ldots, i - 1, i + 1, \ldots, i + 2k \tag{3.18}$$

will not be taken into further consideration.

As mentioned before, a jump between x_i and x_{i+1} will not affect the test statistics related to $\mathrm{H}_0^j, j = i - 2k, \ldots, i - k$ and $j = i + k, \ldots, i + 2k$, so the rule of Wu and Chu (1993) is rather restrictive and we also consider an alternative set $\Psi_{b+1}^{\prime,a}$ in Step 3 by defining $\Psi_{b+1}^{\prime,a} = \Psi_b^{\prime,a} \setminus \{\psi_b - k, \ldots, \psi_b + k\}$. We will denote the estimators of the resulting jump locations by $\widehat{\xi}_1', \ldots, \widehat{\xi}_{\widehat{m}}'$ and call the two procedures WC2 and WC1, depending on the number of subtracted observations $2k$ and $1k$, respectively.

An alternative algorithm was introduced by Qiu (1994) and used for local polynomial jump detection in nonparametric regression by Qiu and Yandell (1998). Assume that Ψ_1^a has cardinality $\mathrm{card}(\Psi_1^a) = \zeta$ and its elements are given by $k \leq \varphi_1 < \ldots < \varphi_\zeta \leq n - k$. We start with the smallest index φ_1 and put indices, which are very close together and possibly belong to the same jumps, in the same subset, until the last index φ_ζ is used. Formally denoted:

If indices $r_{1;\tilde{b}}$ and $r_{2;\tilde{b}}$ exist with $k \leq r_{1;\tilde{b}} \leq r_{2;\tilde{b}} \leq n - k$, such that

$$\begin{aligned} \varphi_{r_{1;\tilde{b}}} - \varphi_{(r_{1;\tilde{b}}-1)} &> 2k \\ \varphi_{(r_{2;\tilde{b}}+1)} - \varphi_{r_{2;\tilde{b}}} &> 2k \\ \varphi_{\kappa+1} - \varphi_\kappa &\leq 2k, \text{ for } r_{1;\tilde{b}} \leq \kappa \leq (r_{2;\tilde{b}} - 1), \end{aligned} \tag{3.19}$$

then $r_{1;\tilde{b}}$ and $r_{2;\tilde{b}}$ give the smallest and the largest index of subset $\tilde{b}$, respectively, with

$$\Phi_{r_{1;\tilde{b}}, r_{2;\tilde{b}}, \tilde{b}} = \{\varphi_{r_{1;\tilde{b}}}, \ldots, \varphi_{r_{2;\tilde{b}}}\} \subseteq \Psi_1^a \text{ and } \widehat{\xi}_{1;\tilde{b}} = \frac{x_{\varphi_{r_{1;\tilde{b}}}} + x_{\varphi_{r_{2;\tilde{b}}}}}{2}, \tilde{b} = 1, \ldots, \tilde{m} \tag{3.20}$$

as the related $\tilde{m}$ estimators of the jump-locations, whereas $r_{1;1} = \varphi_1$ and $r_{2;\tilde{m}} = \varphi_\zeta$. In case that $r_{1;\tilde{b}} = r_{2;\tilde{b}}$ the resulting subset $\Phi_{r_{1;\tilde{b}}, r_{2;\tilde{b}}, \tilde{b}}$ includes only one point and $\widehat{\xi}_{1;\tilde{b}} = x_{\varphi_{r_{1;\tilde{b}}}}$ is the estimated jump location of the subset. We use $\tilde{m}$ as estimated value for m and $\widehat{\xi}_{1;1}, \ldots, \widehat{\xi}_{1;\tilde{m}}$ as estimators for the m jump locations $\xi_1, \ldots, \xi_m$.

Instead of taking the middle point $\widehat{\xi}_{1;\tilde{b}}$ from each subset $\Phi_{r_{1;\tilde{b}},r_{2;\tilde{b}},\tilde{b}}$, as proposed by Qiu and Yandell (1998), we also consider the two alternatives

$$\overline{\xi}_{2;\tilde{b}} = \frac{1}{\mathrm{card}(\Phi_{r_{1;\tilde{b}},r_{2;\tilde{b}},\tilde{b}})} \sum_{j \in \Phi_{r_{1;\tilde{b}},r_{2;\tilde{b}},\tilde{b}}} x_{\varphi_j} \tag{3.21}$$

and

$$\widetilde{\xi}_{3;\tilde{b}} = x_{\varphi^\star_{r_{1;\tilde{b}},r_{2;\tilde{b}}}}, \text{ with } \varphi^\star_{r_{1;\tilde{b}},r_{2;\tilde{b}}} = \underset{j \in \Phi_{r_{1;\tilde{b}},r_{2;\tilde{b}},\tilde{b}}}{\arg\max} |T_a(\Upsilon_j)|, \tag{3.22}$$

which are the average over all candidates from $\Phi_{r_{1;\tilde{b}},r_{2;\tilde{b}},\tilde{b}}$ and the candidate of $\Phi_{r_{1;\tilde{b}},r_{2;\tilde{b}},\tilde{b}}$ with the smallest p-value, respectively. Like we have proposed for the method of Wu and Chu (1993), we also compare the procedure of Qiu and Yandell (1998) in a less restrictive way, by using the distance k instead of $2k$, to find the subsets of indices. We will denote the six different rules with QY11, QY12, QY13, QY21, QY22 and QY23, the first digit gives the minimal length $1k$ or $2k$ between two subsets and the second digit gives the used method to calculate the jump candidate of a given subset.

3.2.4 Detection rules: Robust cross-Validation

In the following we call a detection rule a combination of one of the two sample test statistics $T_a, a = 1, \ldots, 7$ from Section 3.2.1 with a level of significance chosen fixed or by Bonferroni or Bonferroni-Holm and one of the methods to find the jump locations within all candidates (see Section 3.2.3). If we want to detect a jump between x_i and x_{i+1}, $i = k, \ldots, n-k$, then the value of the test statistic and so the decision, whether a jump is found at position x_i, depends on the parameter k. If we assume i.i.d. rvs in both samples, the test efficiency will increase for an increasing sample size k. But if at least one of the two samples contains outliers or observations before and after a jump, what is more probable for an increasing value of k, this assumption is no longer fulfilled and the power of the test can decrease.

One way to choose k adaptively, i.e. based on the data, is cross-validation (CV). In classical leave-one-out-CV for each point (x_i, y_i) the function value $f(x_i)$ is estimated with the information of all observations except (x_i, y_i). The distance between y_i and the $\widehat{f}(x_i)$ is then used to find a good selection of the smoothing parameter k. For the indirect approach we have to exclude the related observation (x_i, y_i) already before the jump locations are determined:

For a given k and a given detection rule we consider the sample without (x_i, y_i). From the remaining $n-1$ observations we estimate the jump-locations $\widehat{\xi}^{i,k}_1, \ldots, \widehat{\xi}^{i,k}_{\widehat{m}}$ and divide the design points and observations into the $\widehat{m}+1$ subsets

$$\begin{aligned} \mathcal{X}^i_r &= \{x_j, j = 1, \ldots, i-1, i+1, \ldots, n : \widehat{\xi}^{i,k}_{r-1} < x_j \leq \widehat{\xi}^{i,k}_r\} \text{ and} \\ \mathcal{Y}^i_r &= \{y_j, j = 1, \ldots, i-1, i+1, \ldots, n : x_j \in \mathcal{X}^i_r\}, r = 1, \ldots, \widehat{m}+1, \end{aligned} \tag{3.23}$$

with $\widehat{\xi}^{i,k}_0 = 0$ and $\widehat{\xi}^{i,k}_{\widehat{m}+1} = 1$ and consider r with $x_i \in \mathcal{X}^i_r$. The function value $f(x_i)$ can then be estimated by $\widehat{f}_{-i;k}(x_i) = \tilde{\Xi}_a(\mathcal{Y}^i_r)$, with $\tilde{\Xi}_a$ being a location estimator related to

the test statistic T_a. For the t-test T_1, $\tilde{\Xi}_1$ is the sample mean, for the trimmed t-test T_2, $\tilde{\Xi}_2$ is the 25%-trimmed-mean, for the median comparison T_3 and the median-test T_7, $\tilde{\Xi}_3 = \tilde{\Xi}_7$ is the sample median and for the two Hodges-Lehmann-tests T_4 and T_5 and the Wilcoxon-rank-sum-test T_6, $\tilde{\Xi}_4 = \tilde{\Xi}_5 = \tilde{\Xi}_6$ is the one-sample HLE. Estimating $\widehat{f}_{-i;k}(x_i)$ for each i gives us the n cross-validated residuals $\widehat{e}_{i;k} = y_i - \widehat{f}_{-i;k}(x_i)$.

We consider five different CV-criteria in the following. We refer to the results of Chapter 2, where several CV-criteria were compared for jump-preserving smoothing. Beside the common L_2-CV

$$LSCV(k) \;=\; \arg\min_k \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{f}_{-i;k}(x_i)\right)^2 \;=\; \arg\min_k \frac{1}{n}\sum_{i=1}^{n}\widehat{e}_{i;k}^{\,2} \tag{3.24}$$

we also include L_1-CV

$$LADCV(k) \;=\; \arg\min_k \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{f}_{-i;k}(x_i)| \;=\; \arg\min_k \frac{1}{n}\sum_{i=1}^{n}|\widehat{e}_{i;k}|, \tag{3.25}$$

as it is often used in connection with the direct approach (Yang and Zheng, 1992; Wang and Scott, 1994), see Section 2.3 for details. Both L_2- and L_1-CV have problems in robust jump-preserving, if large outliers are observed. Therefore it is of interest, how they perform in the indirect approach considered here.

To include an alternative with a highly robust criterion, we also consider median-CV (Zheng and Yang, 1998)

$$MEDCV(k) \;=\; \arg\min_k \operatorname{Med}\left(|\widehat{e}_{1;k}|, |\widehat{e}_{2;k}|, \ldots, |\widehat{e}_{n;k}|\right). \tag{3.26}$$

The last two criteria are taken, because of their good performance in the direct approach in situations with outliers and jumps. We consider one M-CV-criterion

$$M_\rho CV(k) \;=\; \arg\min_k \frac{1}{n}\,\widehat{\sigma}_k^2 \sum_{i=1}^{n}\rho\left(\frac{\widehat{e}_{i;k}}{\widehat{\sigma}_k}\right) \;. \tag{3.27}$$

by using the redescending Tukey-ρ-function

$$\rho_T(z) = \left(1 - \left(1 - \left(\frac{z}{4.685}\right)^2\right)^3\right)\mathbb{1}_{[0,4.685]}(|z|) + \mathbb{1}_{(4.685,\infty)}(|z|) \tag{3.28}$$

and the approach of Bianco and Boente (2006) and Boente and Rodriguez (2008)

$$BOECV(k) \;=\; \arg\min_k \operatorname{Med}^2\left(\widehat{e}_{1;k}, \ldots, \widehat{e}_{n;k}\right) \;+\; Q_n^2\left(\widehat{e}_{1;k}, \ldots, \widehat{e}_{n;k}\right), \tag{3.29}$$

which we will call again Boente-CV.

3.3 Comparison of the detection rules

3.3.1 Measuring the quality of a jump detection rule

In the following we introduce some measures of accuracy for estimating the true number of jumps m, the jump locations $\xi_1, \ldots, \xi_m$ and the true function f. Imagine we have ν replications of the same data situation and each replication consists of n observations. For the interesting quantities and a given jump detection rule we obtain

$$\widehat{m}^\lambda, \quad \widehat{\xi}_1^\lambda, \ldots, \widehat{\xi}_{\widehat{m}^\lambda}^\lambda \quad \text{and} \quad \widehat{f}(x_i^\lambda), \quad \lambda = 1, \ldots, \nu, \ i = 1, \ldots, n \tag{3.30}$$

as estimations. We further denote

$$\begin{aligned} \mathcal{X}_r &= \{x_i, i = 1, \ldots, n : \widehat{\xi}_{r-1} < x_i \leq \widehat{\xi}_r\} \quad \text{and} \\ \mathcal{Y}_r &= \{y_i, i = 1, \ldots, n : x_i \in \mathcal{X}_r\}, \quad r = 1, \ldots, \widehat{m}^\lambda + 1, \end{aligned} \tag{3.31}$$

with $\widehat{\xi}_0 = 0$ and $\widehat{\xi}_{\widehat{m}^\lambda+1} = 1$, as the subsets that include all design points and the corresponding observations, respectively, belonging to the design interval between two estimated subsequent jumps. The estimation of f at position x_i is done again with a location estimator $\tilde{\Xi}_a, a = 1, \ldots, 7$, applied to that $\mathcal{Y}_r$ with $\widehat{\xi}_{r-1} < x_i \leq \widehat{\xi}_r$.

For the true number of jumps m, two things are of interest here. Firstly, how often is m estimated correctly and secondly, how far are the estimations away from m on average . Let $\widehat{m}^\lambda$ be the estimated number of jumps for the λ-th data set, then

$$\overline{\Delta}_1 = \frac{1}{\nu} \sum_{\lambda=1}^{\nu} \mathbb{1}_{\{0\}}(|\widehat{m}^\lambda - m|) \tag{3.32}$$

is the relative frequency that the estimated number of jumps is equal to m and

$$\overline{\Delta}_2 = \frac{1}{\nu} \sum_{\lambda=1}^{\nu} |\widehat{m}^\lambda - m| \tag{3.33}$$

measures the absolute mean distance between true and estimated number of jumps. It is reasonable to consider both measures, as we are generally interested in estimating the true m correctly as often as possible (what is measured by $\overline{\Delta}_1$), but without taking a too large loss, if m and $\widehat{m}^\lambda$ are different (what will be measured by $\overline{\Delta}_2$).

On the other hand we are interested in a good estimation of the true jump locations $\xi_1, \ldots, \xi_m$. Again two issues are of interest here. Firstly, we want our estimated jump locations to be as close as possible to the true ones. Secondly, we want as few as possible wrongly detected jump locations. For the first question we consider

$$\check{\Delta}^\lambda = \frac{1}{m} \sum_{\iota=1}^{m} \check{\Delta}_\iota^\lambda, \ \text{with} \ \check{\Delta}_\iota^\lambda = \mathbb{1}_{\{1,\ldots,\widehat{m}^\lambda\}} \left(\sum_{\iota'=1}^{\widehat{m}^\lambda} \mathbb{1}_{J_\iota^\lambda}(\widehat{\xi}_{\iota'}^\lambda - \xi_\iota) \right) \tag{3.34}$$

with the interval

$$J_\iota^\lambda = [\min(\xi_\iota - 2\delta^\lambda, x_{\xi_\iota;-}^\lambda), \max(\xi_\iota + 2\delta^\lambda, x_{\xi_\iota;+}^\lambda)], \tag{3.35}$$

which includes the average distance between two neighboured design points

$$\delta^\lambda = \frac{1}{n-1} \sum_{i=1}^{n-1} \left(x_{i+1}^\lambda - x_i^\lambda \right) \tag{3.36}$$

as well as $x^\lambda_{\xi_\iota;-}$ and $x^\lambda_{\xi_\iota;+}$, which are the largest design point smaller than the jump ξ_ι and the smallest design point larger than ξ_ι, respectively. Preliminary calculations show that using $[\xi_\iota - \delta^\lambda, \xi_\iota + \delta^\lambda]$ deliver a too hard criterion, while $[\xi_\iota - 2\delta^\lambda, \xi_\iota + 2\delta^\lambda]$ gives better possibilities for comparison. As we work with a random design it is possible that no design point is in the latter interval. To prevent this case, we choose J_ι^λ from (3.35), what allows at least one design point before and after ξ_ι, respectively, to be detected as jump candidate without taking the risk to obtain a too short interval.

The value $\check{\Delta}^\lambda$ measures the relative frequency of the detected jumps in sample λ. A jump ξ_ι is regarded as detected, if at least one of the estimated jump locations is within J_ι^λ. The average value

$$\overline{\Delta}_3 = \frac{1}{\nu} \sum_{\lambda=1}^{\nu} \check{\Delta}^\lambda = \frac{1}{\nu m} \sum_{\lambda=1}^{\nu} \sum_{\iota=1}^{m} \check{\Delta}_\iota^\lambda \tag{3.37}$$

gives the mean relative frequency of the detected jumps over all ν data sets. If $\widehat{m}^\lambda = 0$ we do not detect any jump in this data set and to that effect is $\check{\Delta}^\lambda = 0$.

Detection rules, which lead to similar values of $\overline{\Delta}_3$, may differ in the number of estimated jump locations, which are too far away from any true jump. As we also want to prevent incorrectly detected jump locations, we consider an alternative measure, which penalises estimated values $\widehat{\xi}_{\iota'}^\lambda$, which have too large distances to all true jump locations, namely

$$\overline{\Delta}_4 = \frac{1}{\hat{\nu}} \sum_{\lambda=1}^{\nu} \hat{\Delta}^\lambda \quad \text{with} \quad \hat{\Delta}^\lambda = \max_{\iota' \in \{1,\ldots,\widehat{m}^\lambda\}} \min_{\iota \in \{1,\ldots,m\}} |\widehat{\xi}_{\iota'}^\lambda - \xi_\iota| \ . \tag{3.38}$$

$\overline{\Delta}_4$ calculates the averaged distance of the estimated jump location, which has the largest distance to its closest ξ_ι. The max-min value is advantageous over an simple average, as we are especially interested in those jump locations, which do not belong to any ξ_ι, as they have a too large distance to any true jump location.

The last question is how good the true f is fitted, after a detection rule with subsequent estimation $\widehat{f}$ is performed. The performance of $\widehat{f}$ for a single data set is measured by the Averaged Squared Error (ASE; see Haerdle 2002, pp. 90)

$$\tilde{\Delta}_A^\lambda = \frac{1}{n} \sum_{i=1}^{n} \left(\widehat{f}(x_i^\lambda) - f(x_i^\lambda) \right)^2 . \tag{3.39}$$

and for all ν data sets by the mean ASE-value (MASE)

$$\overline{\Delta}_A = \frac{1}{\nu} \sum_{\lambda=1}^{\nu} \tilde{\Delta}_A^\lambda = \frac{1}{\nu n} \sum_{\lambda=1}^{\nu} \sum_{i=1}^{n} \left(\widehat{f}(x_i^\lambda) - f(x_i^\lambda) \right)^2 \tag{3.40}$$

For the following simulations we take $n = 200$ and $\nu = 100$ for all data situations.

3.3.2 Planning and evaluating the simulation design

We consider model (3.1) with $X_1, \ldots, X_n$ being drawn from an uniform random design, i.e. $X_1, \ldots, X_n$ being i.i.d. uniformly distributed on the interval $[0, 1]$. The noise terms $E_1, \ldots, E_n$ are i.i.d. $\mathcal{N}(0, 1)$ with some outliers at positions chosen at random. For obtaining π percent outliers $\max\{\lfloor n\pi \rfloor, 1\}$ of all n positions are drawn without replacement. At outlier positions, the value $\pm\gamma$ is added to the noise, with the same sign as the closest level shift to produce a more challenging situation for the tests and so for the jump detection rules. The regression function f is chosen as piecewise constant with m jumps, each of height s. The positions $\xi_1, \ldots, \xi_m$ of the discontinuities within the interval (0,1) are fixed.

As we have seen in the Sections before, there are many options in constructing a jump detection rule. If we only consider the methods described here, we have

(1) seven test procedures to find possible jump candidates: $T_1, \ldots, T_7$
(2) three ways to choose the level of significance: α_f, α_b and α_h
(3) eight rules to select jump candidates: the method of Wu and Chu and three variants of the method of Qiu and Yandell, each with a distance k and $2k$
(4) five CV-criteria to select k: L_2-, L_1-, median-, Tukey- and Boente-CV.

implying $q = 840$ possible jump detection rules altogether.

To compare these 840 detection rules for p different data situations, we do the comparison in two steps. Firstly we compare for each combination of test statistic and CV-criteria, the performance of the several choices of α and the jump candidate selectors. After finding a recommendation of (2) and (3) for each cross-validated test statistic, we compare the remaining 35 detection rules in a second step.

These two comparisons will be made for some sets of data situations jointly. Each set consists of nine different data situations, as we vary $m \in \{1, 2, 5\}$ and $s \in \{1, 3, 6\}$. For $m = 1$ the jump is located at $\xi_1 = 0.4$, for $m = 2$ the second jump is located at $\xi_2 = 0.6$ and for $m = 5$ we fix $\xi_1 = 0.2, \xi_2 = 0.4, \xi_3 = 0.55, \xi_4 = 0.7$, and $\xi_5 = 0.85$. The sets differ in the percentages and magnitudes of the outliers as we consider the cases (π, γ) with $(0.01, 3)$, $(0.05, 12)$, $(0.15, 12)$, $(0.30, 12)$, $(0.15, 48)$ and $(0.15, 192)$.

In order to simplify the evaluation for a set, we define a summary measure to compare the relative performance of the detection rules. For the η-th data situation and a given accuracy measure $\overline{\Delta}$, we consider the relative loss

$$\Lambda^{\upsilon}_{\eta;1} = \frac{\overline{\Delta}^{\upsilon}_{\eta} - \overline{\Delta}^{\star}_{\eta;1}}{\overline{\Delta}^{\star}_{\eta;1}} \quad \text{and} \quad \Lambda^{\upsilon}_{\eta;2} = \overline{\Delta}^{\star}_{\eta;2} - \overline{\Delta}^{\upsilon}_{\eta} \,, \tag{3.41}$$

respectively, depending on the measure $\overline{\Delta}$. If small values for $\overline{\Delta}$ are desirable, like for $\overline{\Delta}_2, \overline{\Delta}_4$ and $\overline{\Delta}_5$, $\Lambda^{\upsilon}_{\eta;1}$ is the relative loss, if large values for $\overline{\Delta}$ are desirable, like for $\overline{\Delta}_1$ and $\overline{\Delta}_3$, $\Lambda^{\upsilon}_{\eta;2}$ is the relative loss. $\overline{\Delta}^{\upsilon}_{\eta}$ is the realised value of $\overline{\Delta}$ for data situation η and estimator υ, $\eta = 1, \ldots, p$ and $\upsilon = 1, \ldots, q$. The values

$$\overline{\Delta}^{\star}_{\eta;1} = \min\left(\overline{\Delta}^{1}_{\eta}, \ldots, \overline{\Delta}^{q}_{\eta}\right) \quad \text{and} \quad \overline{\Delta}^{\star}_{\eta;2} = \max\left(\overline{\Delta}^{1}_{\eta}, \ldots, \overline{\Delta}^{q}_{\eta}\right) \tag{3.42}$$

give the criterion value of the best estimator for the given data situation η, depending on $\overline{\Delta}$. Values of Λ_η^υ close to zero indicate that detection rule υ performs almost as well as the best detection rule for situation η. We use the mean relative loss (MRL)

$$\overline{\Lambda}_1^\upsilon = \frac{1}{p}\sum_{\eta=1}^{p}\Lambda_{\eta;1}^\upsilon \quad \text{and} \quad \overline{\Lambda}_2^\upsilon = \frac{1}{p}\sum_{\eta=1}^{p}\Lambda_{\eta;2}^\upsilon \tag{3.43}$$

as global performance measure in the comparison for the included data situations for the accuracy measures $\overline{\Delta}_c, c = 1, \ldots, 4, A$ for the different sets of data situations.

3.3.3 Choosing the level of significance and the selection criterion for the jump candidates

In a first step we will analyse the performance of the different combinations of choosing the level of significance α and selecting good jump candidates for the different cross-validated test procedures. We call the alternatives of choosing the level of significance α-selectors and the procedures of Wu and Chu (1993) and Qiu and Yandell (1998) jump-selectors in the following. For this first comparison, the MRL (3.43) based on all 54 data situations of the six sets described in Section 3.3.2 will be calculated for each accuracy measure $\overline{\Delta}_c, c = 1, \ldots, 4, A$. For each of these measures, we will only present the graphic of one CV-criterion for four different tests and comment shortly the results for the other cross-validated test statistics. The following results show

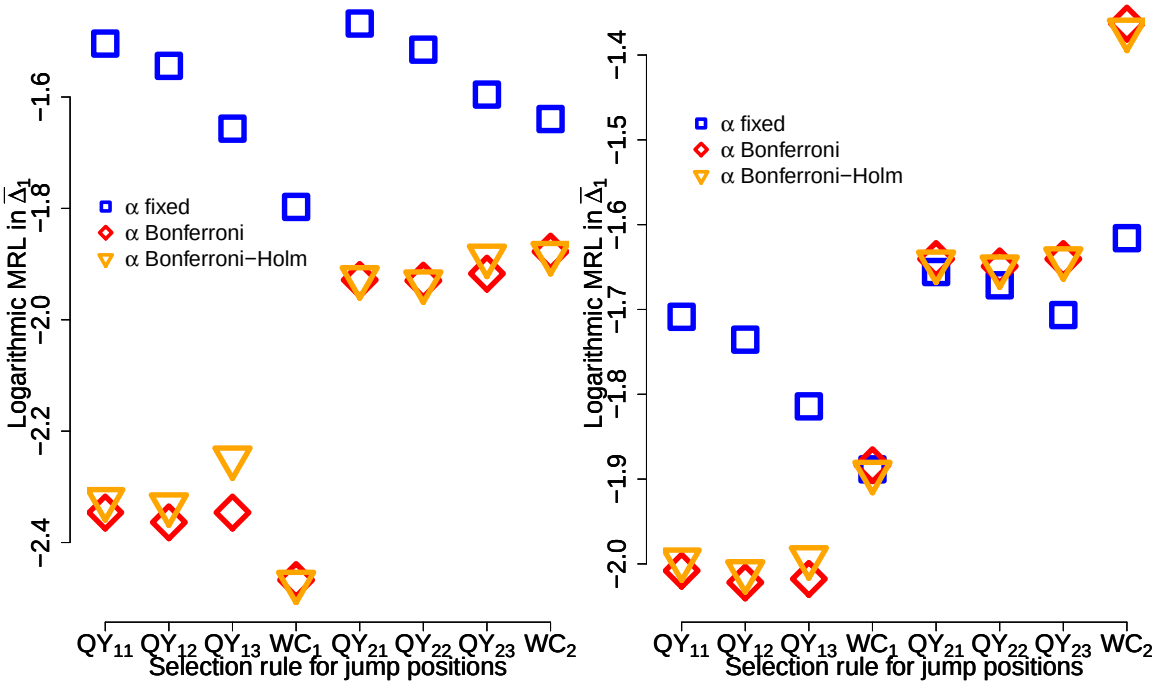

Figure 3.1: MRL in $\overline{\Delta}_1$ for the trimmed t-test (left) and the Wilcoxon-test (right), each with Tukey-CV.

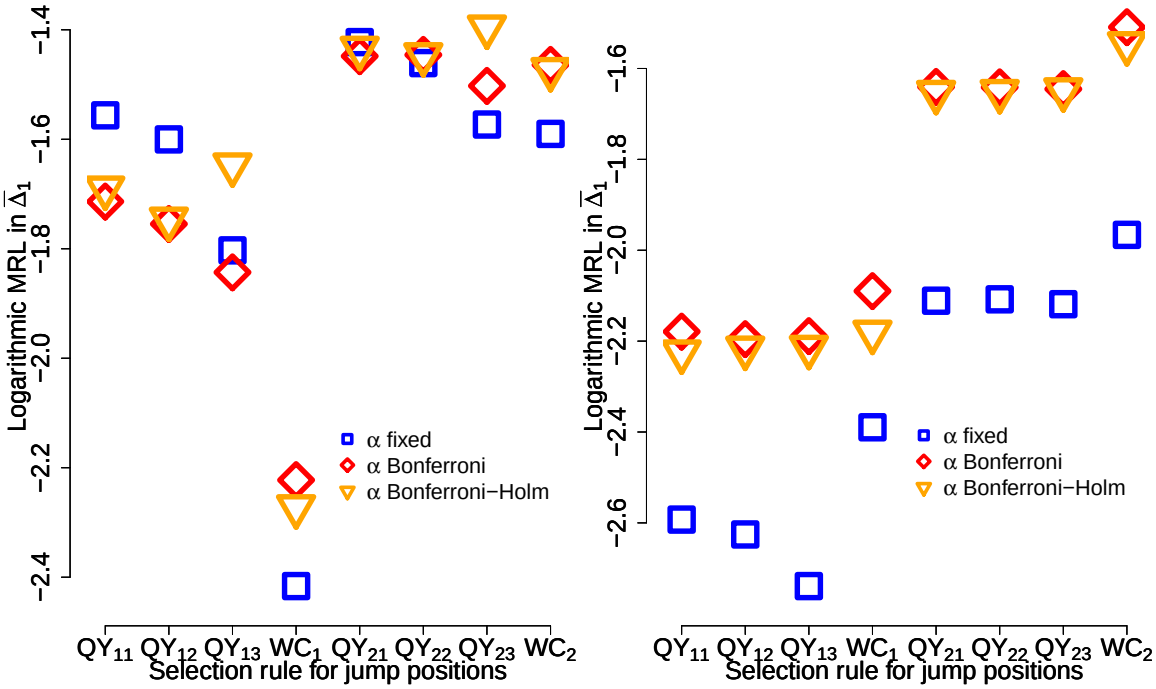

Figure 3.2: MRL in $\overline{\Delta}_1$ for the tests based on HLEs T_4 (left) and T_5 (right), each with Tukey-CV.

that for the different criteria $\overline{\Delta}_c$ we gain different recommendations of α- and jump-selectors. Note that we use a logarithmic scale for the ordinate, because we want to visualise differences among the better detection rules.

For the relative proportion of data sets Δ_1 with a correctly estimated number of jumps m, a procedure with a smaller distance $1k$ is preferable. For the t-test, the Wilcoxon-test and the two-sample-HLE test, the rule of Qiu and Yandell is preferable, while the approach of Wu and Chu is better for the trimmed t-test, the median-test and the two robust tests based on the comparisons of one-sample medians and HLEs. An adjusted α delivers smaller losses for the t-tests and the rank tests, but for the three tests of Fried and Dehling, the fixed choice of α is appropriate. Compare Fig. 3.1 and Fig. 3.2 for the results of four tests, each combined with Tukey-CV.

Considering the mean absolute distance Δ_2 between true and estimated number of jumps, the t-tests and the rank tests perform best, if one of the procedures of Qiu and Yandell with the smaller choice $1k$ and an adjusted α is used. See Fig. 3.3 for one cross-validated trimmed t-test and Wilcoxon-test, respectively. For the three tests of Fried and Dehling, the procedure of Wu and Chu, again with the smaller distance $1k$ and α selected by the Bonferroni- or Bonferroni-Holm-method is better, see Fig. 3.4.

For the relative proportion Δ_3 of detected true jump locations, all cross-validated tests have smaller loss values, if the fixed chosen $\alpha = 0.01$ is considered as α-selector. However, for each test another jump-selector seems to be advantageous, but the differences in the MRL-values are rather small. So it seems not to be crucial for Δ_3, which jump-selector to use. See Fig. 3.5 for two cross-validated linear rank tests and 3.6 for two cross-validated tests of Fried and Dehling.

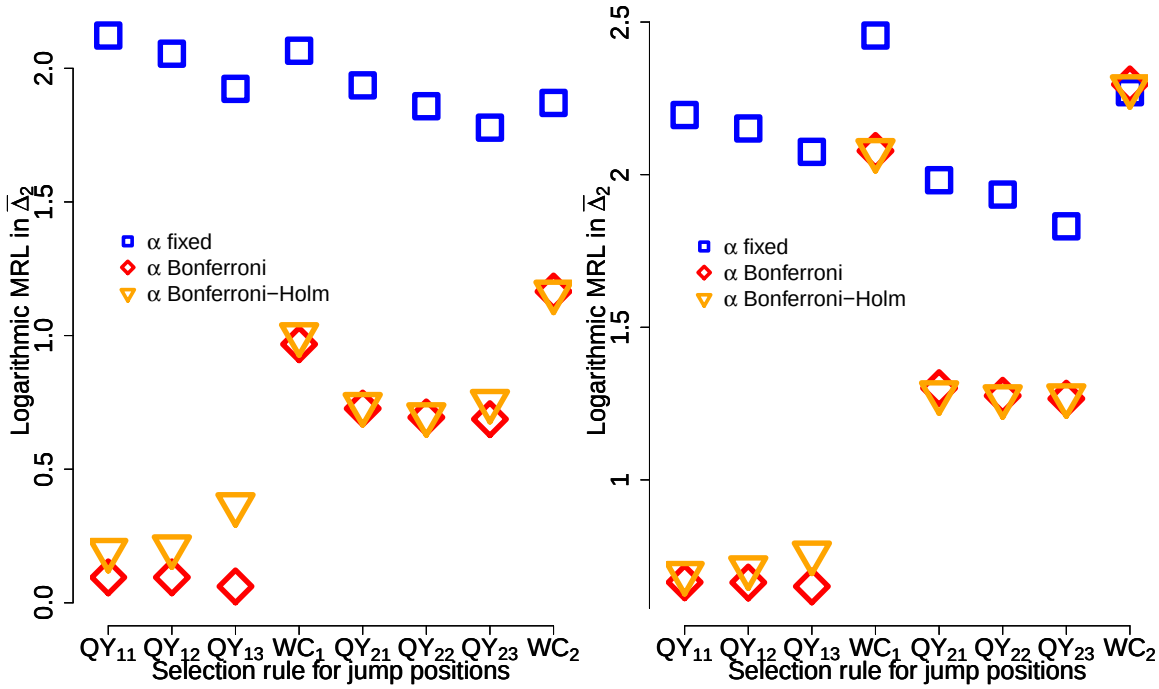

Figure 3.3: MRL in $\overline{\Delta}_2$ for the trimmed t-test (left) and the Wilcoxon-test (right), each with Tukey-CV.

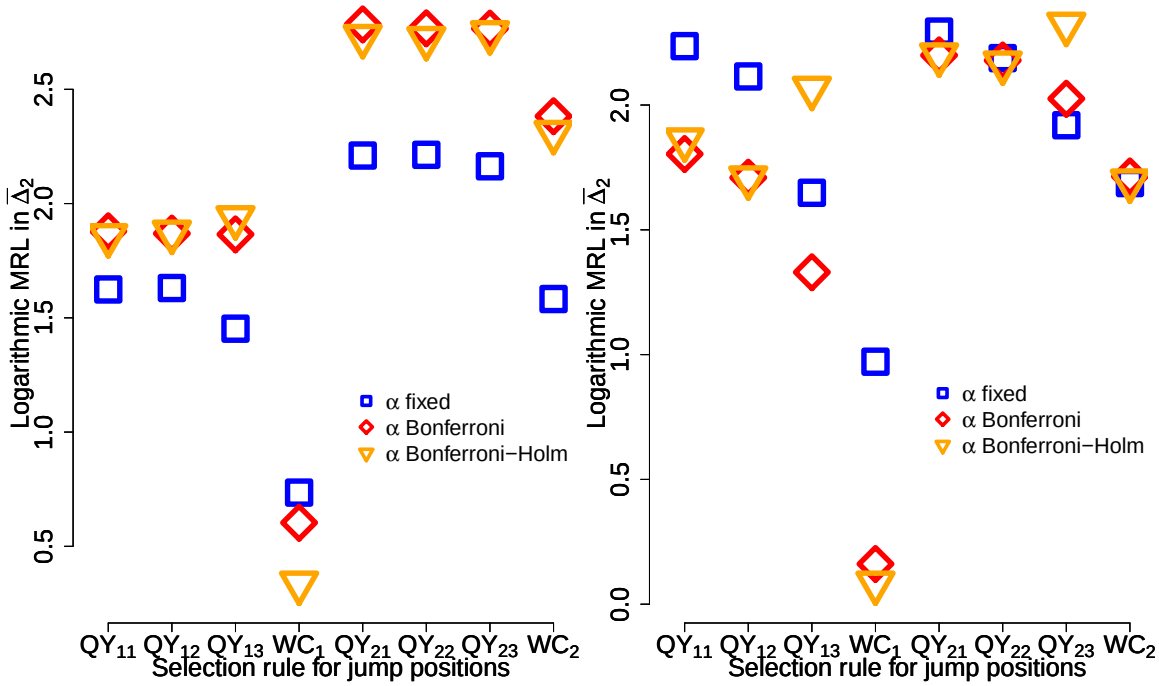

Figure 3.4: MRL in $\overline{\Delta}_2$ for the tests based on differences of medians T_3 (left) and HLEs T_4 (right), each with Tukey-CV.

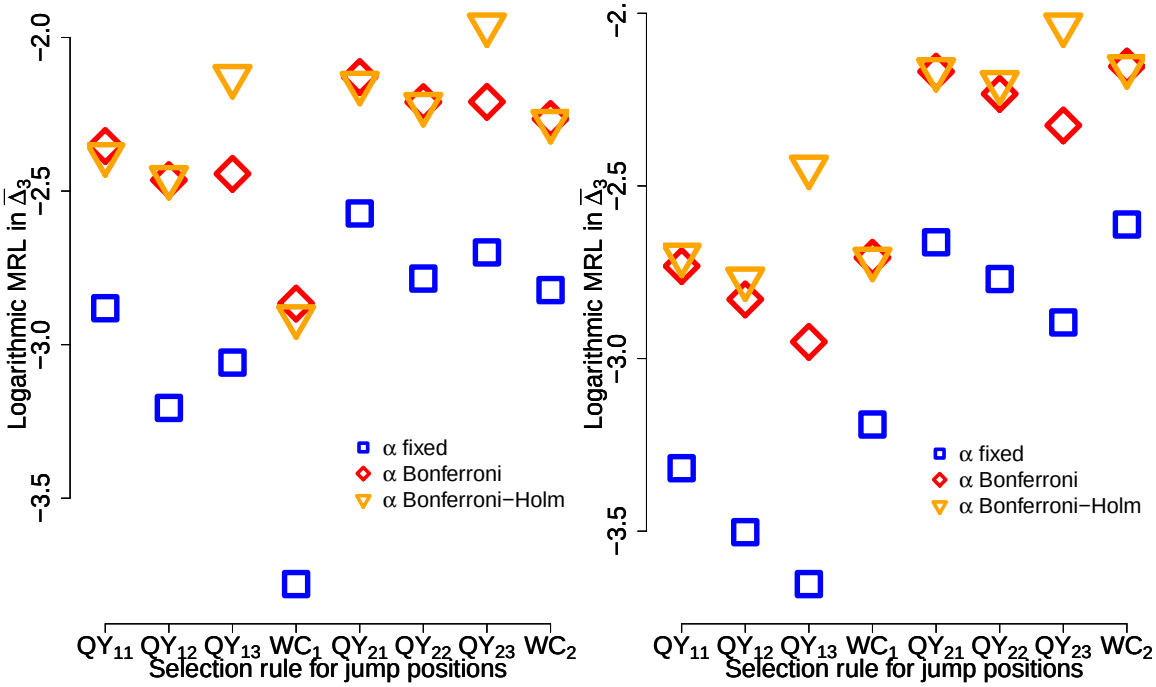

Figure 3.5: MRL in $\overline{\Delta}_3$ for the Wilcoxon-test (left) and the median-test (right), each with Tukey-CV.

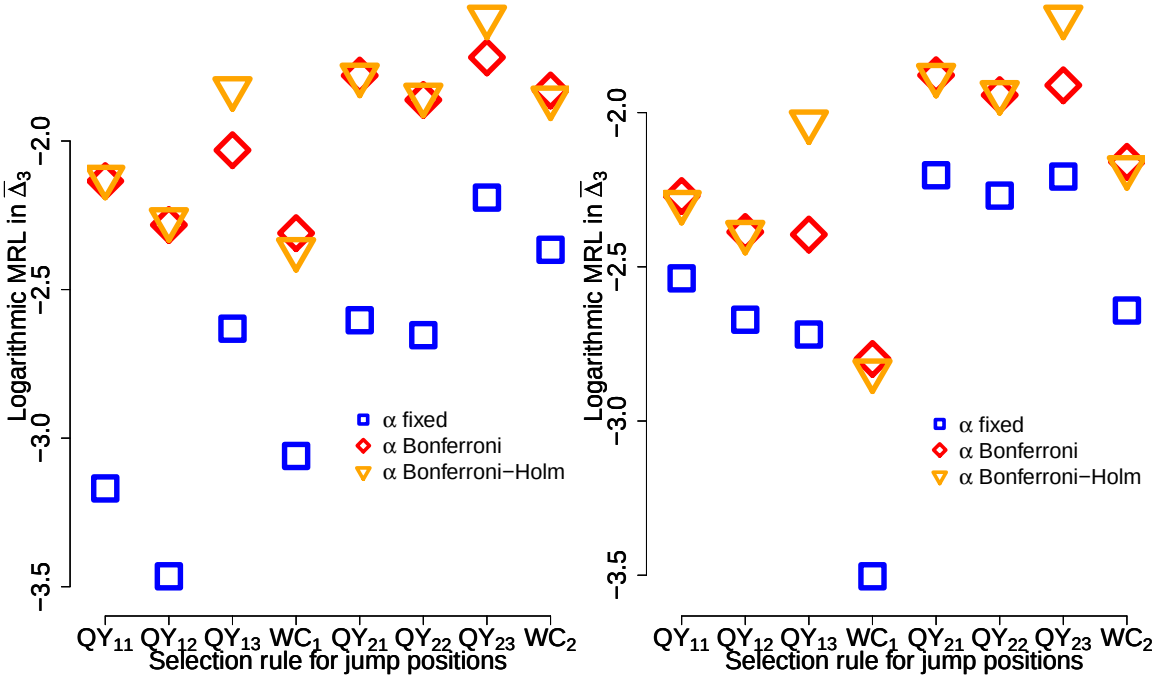

Figure 3.6: MRL in $\overline{\Delta}_3$ for the tests based on differences of medians T_3 (left) and HLEs T_4 (right), each with Tukey-CV.

Analysing the maximal distance Δ_4 of an estimated jump location to a true location ξ_ι, a fixed value of α performs rather poorly for all tests. The criterion QY23, which uses the smallest p-value of a subset and a distance of at least $2k$ points between two subsets, delivers one of the smallest MRL-values for all test-CV-combinations (see Fig. 3.7 for two examples), except the three tests of Fried and Dehling, where again the jump-selector WC1 delivers the smallest MRL, see Fig. 3.8.

Even if the primary aim of the indirect approach is a good estimation of the jump locations, it is also of interest to compare the resulting estimations of f via the MASE $\overline{\Delta}_A$. The results of four tests combined with Tukey-CV are shown in Fig. 3.9 and 3.10. The jump-selectors based on the rule of Wu and Chu perform best for all tests. WC1 delivers the smallest loss for all tests, except the two t-tests, where WC2 is superior. Again a multiple adjusted level α delivers smaller losses than the fixed chosen α.

A comparison of the MRL, calculated over all measures $\Delta_c, c = 1, \ldots, 4, A$, brings the following recommendations: While we observe a smaller loss for the t-tests and the linear rank tests, if the procedure of Qiu and Yandell (1998) is used, the three robust tests of Fried and Dehling (2011) show a better performance with the method of Wu and Chu (1993). For Qiu and Yandell's approach, we observe smaller losses for our modifications in (3.21) and (3.22) than for the original proposal (3.20) of Qiu and Yandell in most of the data situations. The proposal with the smallest p-value (3.22) performs slightly better. For the t-tests, a larger distance $2k$ is preferable, while for the other tests the distance $1k$ is used. This means that for the t-tests we use QY23, for the tests T_3, T_4 and T_5 we use WC1 and for the linear rank tests we use QY13

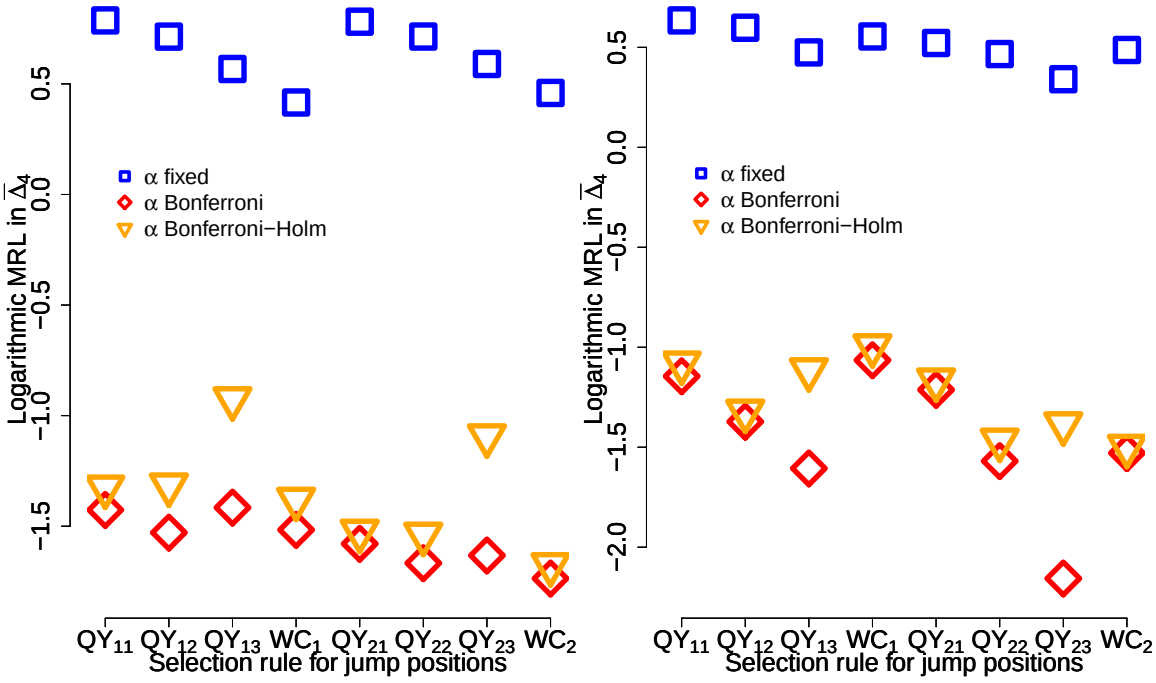

Figure 3.7: MRL in $\overline{\Delta}_4$ for the trimmed t-test (left) and the median-test (right), each with Tukey-CV.

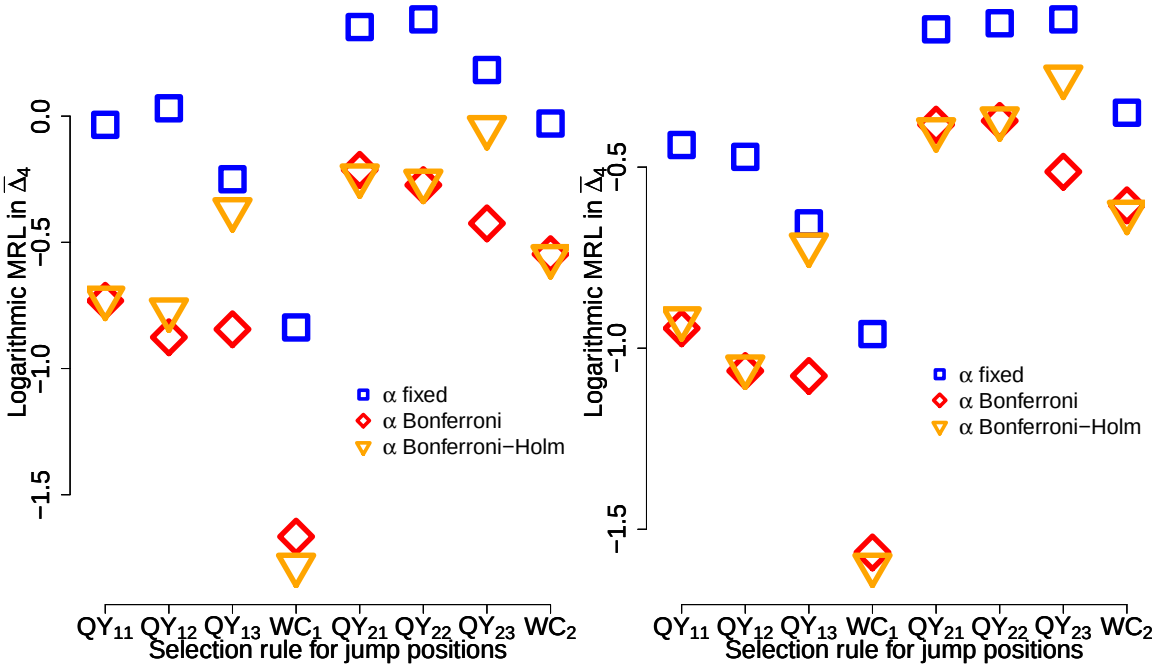

Figure 3.8: MRL in $\overline{\Delta}_4$ for the tests based on differences of medians T_3 (left) and based on the median difference T_5 (right), each with Tukey-CV.

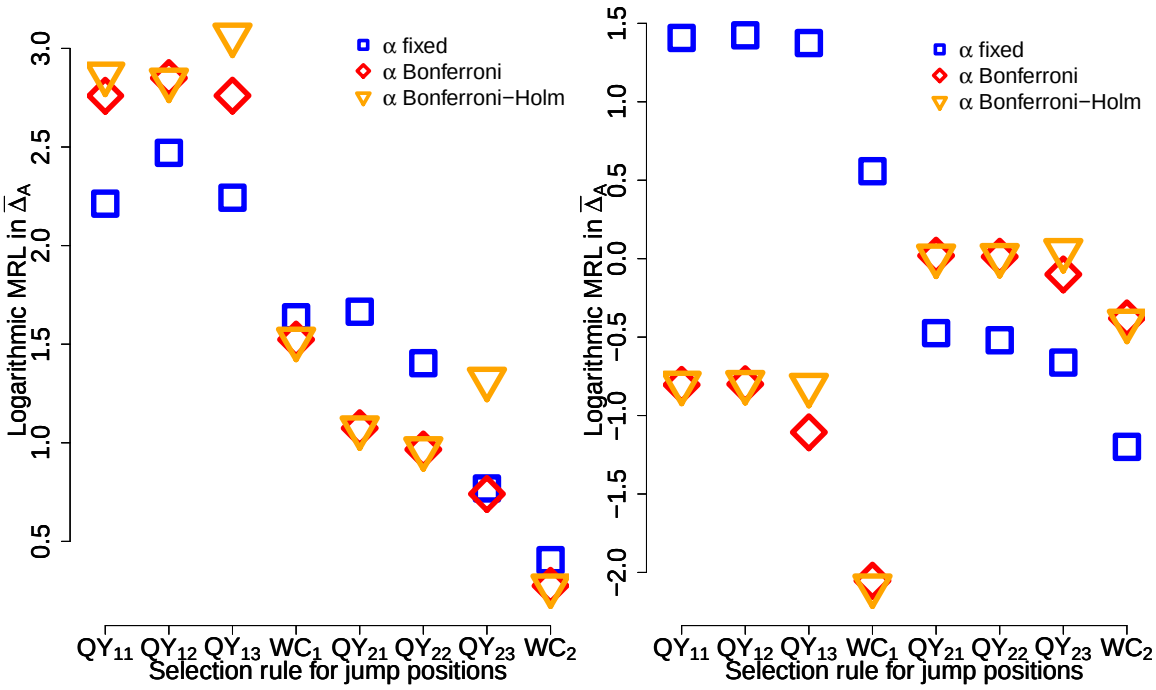

Figure 3.9: MRL in $\overline{\Delta}_A$ for the trimmed t-test (left) and the median-test (right), each with Tukey-CV.

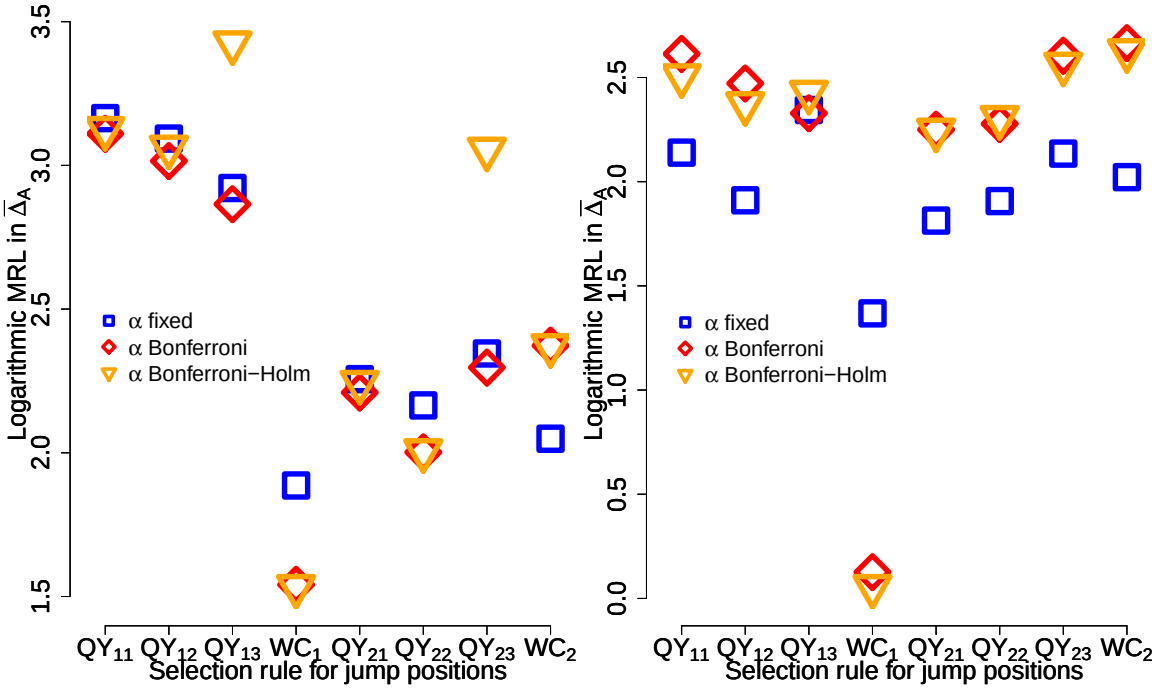

Figure 3.10: MRL in $\overline{\Delta}_A$ for the tests based on HLEs T_4 (left) and T_5 (right), each with Tukey-CV.

for the following analysis. For all tests we use an adjusted level of significance, for the t-tests and the linear rank tests the method of Bonferroni is superior, while for the three tests of Fried and Dehling the procedure of Bonferroni and Holm performs slightly better. With these presettings we will continue in the following.

3.3.4 Comparison of the cross-validated test statistics

This Section includes the final comparison of the cross-validated test statistics with the previous setted α- and jump-selectors from Section 3.3.3. The estimations of the true number of jumps m, the true locations $\xi_1, \ldots, \xi_m$ and the true f are compared.

To compare the estimations of m of the different cross-validated test statistics, the accuracy measures $\overline{\Delta}_1$ and $\overline{\Delta}_2$ are used. Fig. 3.11 shows the MRL-values of the set (0.01,3) without large outliers. While the Wilcoxon-test performs best in terms of the relative proportion of a correctly estimated m, see Fig. 3.11 (left), the two tests based on HLEs deliver smaller loss values in terms of the mean absolute distance of $\widehat{m}$ to m, see Fig. 3.11 (right). Tukey-CV seems to be an appropriate choice in both cases. The different MRL-values for $\overline{\Delta}_1$ and $\overline{\Delta}_2$ are due to the situations with a small jump height $s = 1$ and more than one jump. Here the Wilcoxon-test is the only test, which gains a detection rate over 50% and 5% of the true $m = 2$ and $m = 5$, respectively, leading its much smaller loss values for $\overline{\Delta}_1$. All other tests underestimate m here. However, if the true m is missed, the two tests based on HLEs are most reliable to find an estimation close to m, especially in situations with a larger jump height.

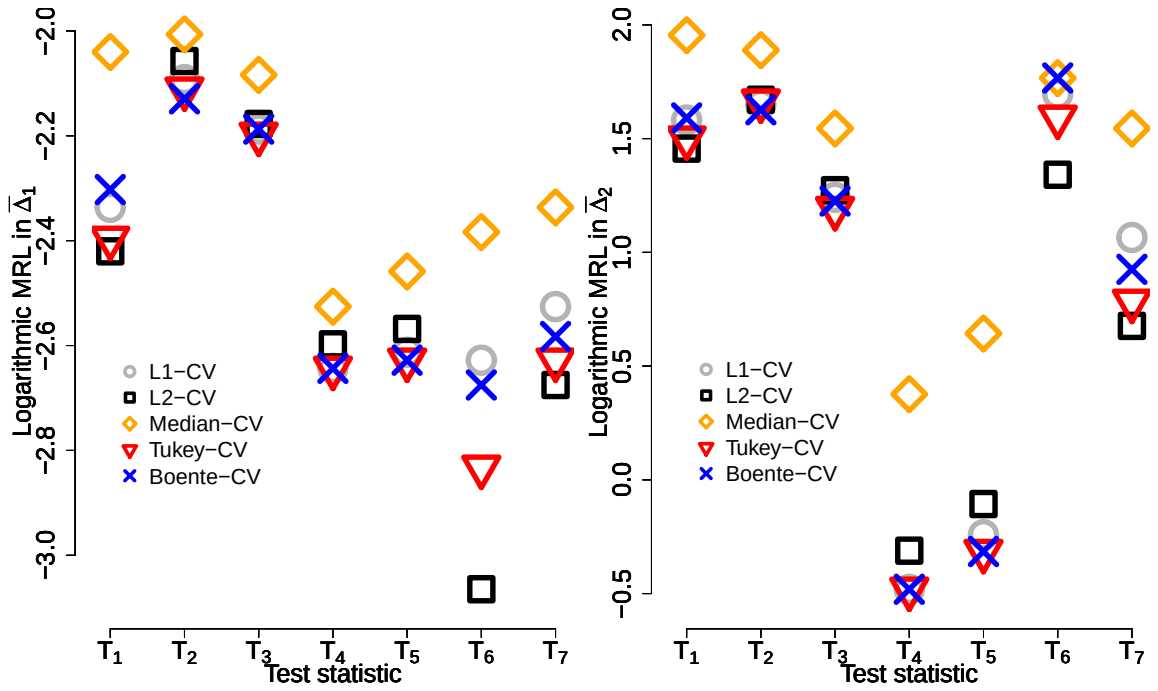

Figure 3.11: MRL in $\overline{\Delta}_1$ (left) and $\overline{\Delta}_2$ (right) for the situation set $(\pi = 0.01, \gamma = 3)$.

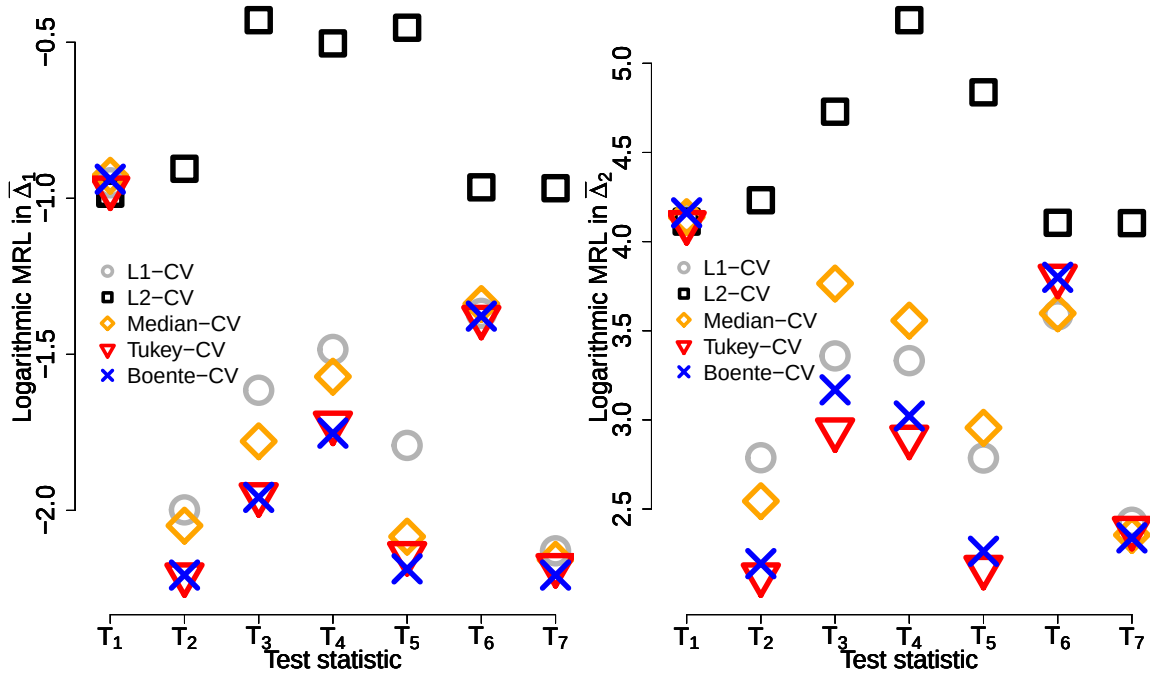

Figure 3.12: MRL in $\overline{\Delta}_1$ (left) and $\overline{\Delta}_2$ (right) for the situation set $(\pi = 0.15, \gamma = 192)$.

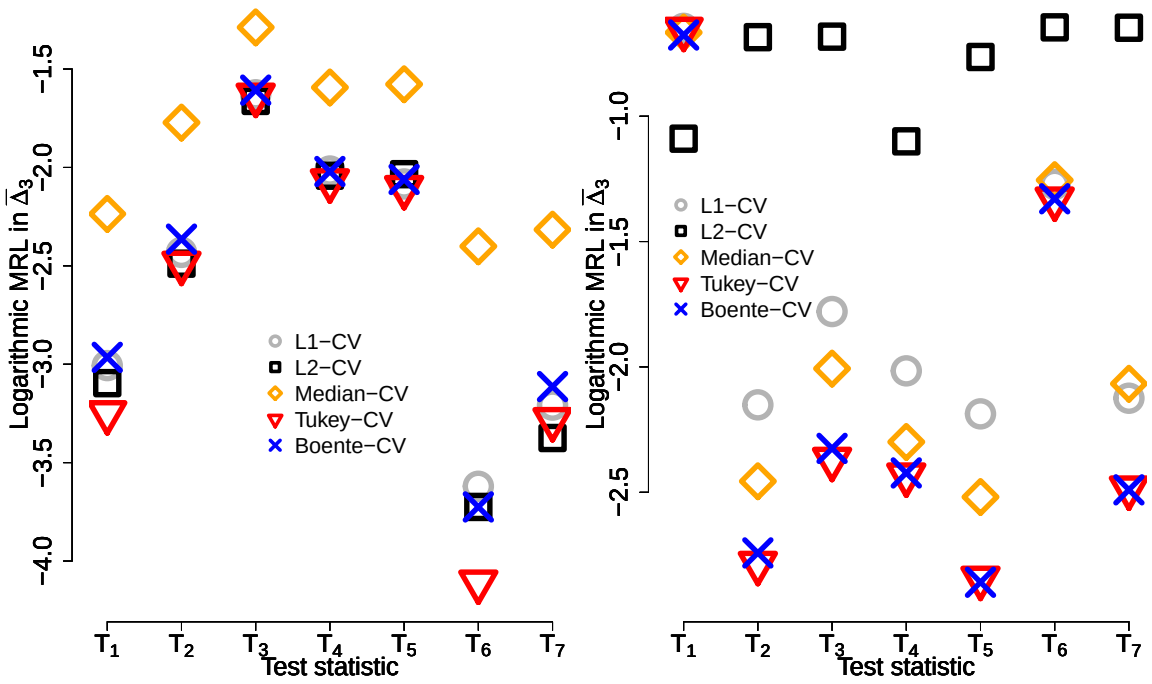

Figure 3.13: MRL in $\overline{\Delta}_3$ for the situation sets $(0.01, 3)$ (left) and $(0.15, 192)$ (right).

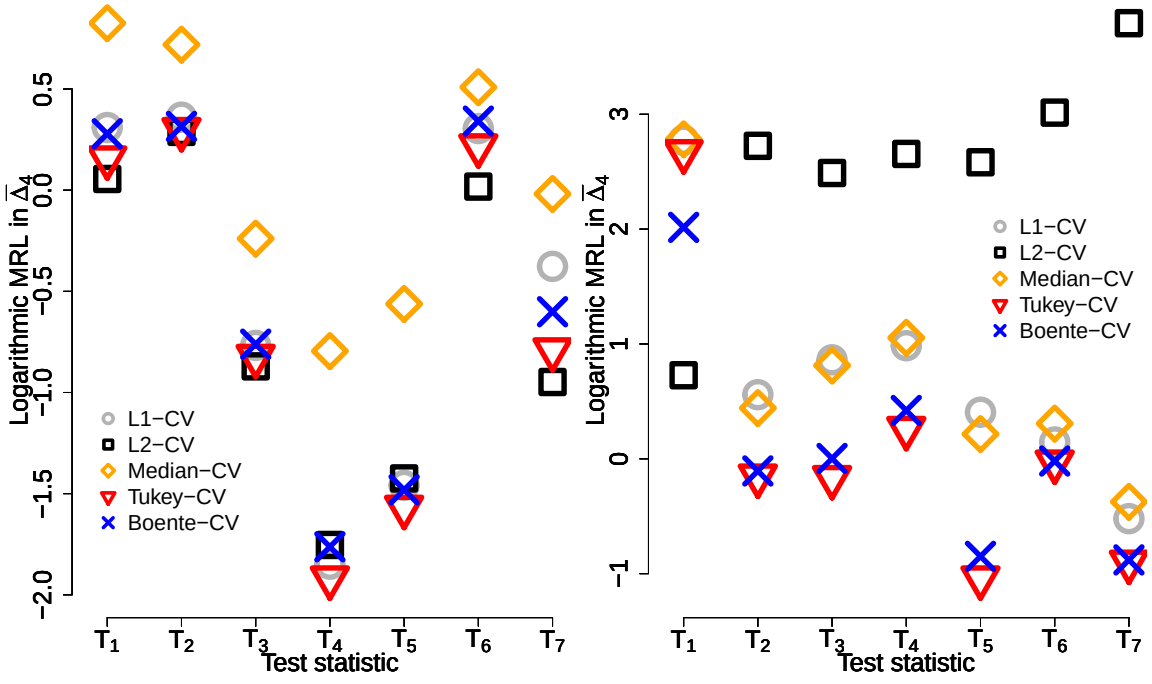

Figure 3.14: MRL in $\overline{\Delta}_4$ for the situation sets $(0.01, 3)$ (left) and $(0.15, 192)$ (right).

For large outliers, see e.g. the results for the set (0.15,192) in Fig. 3.12, the trimmed t-test, the test based on the two-sample-HLE-and the median-test perform best. Tukey- and Boente-CV are superior here, while the unrobust L_2-CV performs poorly in the presence of large outliers. Due to their loss-values, the t-test and the Wilcoxon-test do not give reliable estimations of m, if large outliers are observed.

Looking at the relative proportion $\overline{\Delta}_3$ of the detected jumps, we find in the situation without large outliers, see Fig. 3.13 (left), smaller loss values, when the linear rank-tests or the t-test is used. When large outliers are included the same tests than in 3.12 are preferable. Altogether the median test seems to be a proper choice, if outliers with an arbitrary size are observed and the primary objective is the detection of the true jumps. Furthermore, this test is also recommendable for larger percentages of contamination (not shown), as it is only outperformed by the median comparison T_3 for the situation set (30,12). The use of Tukey-CV is again recommendable, as this criterion delivers good results for all good tests for the respective data situation.

In terms of $\overline{\Delta}_4$, which measures the averaged maximum distance of an estimated jump location to its closest true jump, we find that the two Hodges-Lehmann tests give the smallest loss values, see Fig. 3.14 (left). For large outliers the results can be found in Fig. 3.14 (right). Here, the two-sample-Hodges-Lehmann test performs best. Note that Tukey-CV is again the proper choice for all robust tests.

In the last step we compare the estimations of the true function f. Thereby it has to be considered that estimations do not only depend on the jump detection rule, but also on the corresponding location estimator $\tilde{\Xi}_a, a = 1, \ldots, 7$. In terms of the MASE, the jump detection rules based on the t-test combined with L_2-CV perform

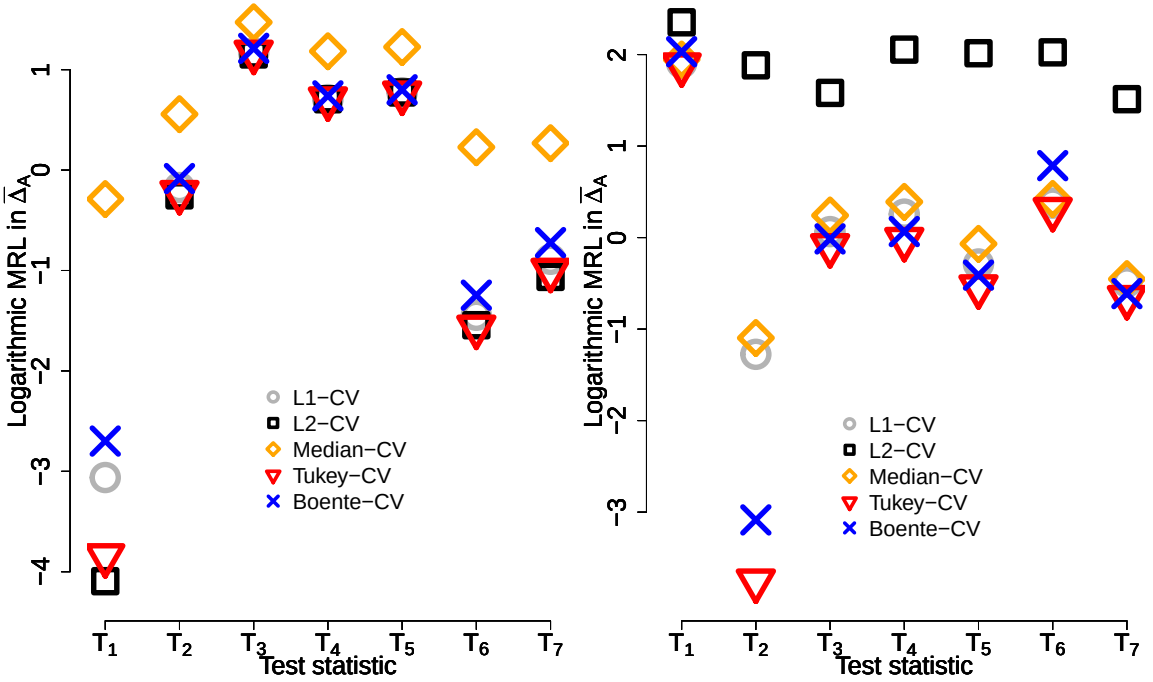

Figure 3.15: MRL in $\overline{\Delta}_A$ for the situation sets $(0.01, 3)$ (left) and $(0.15, 12)$ (right).

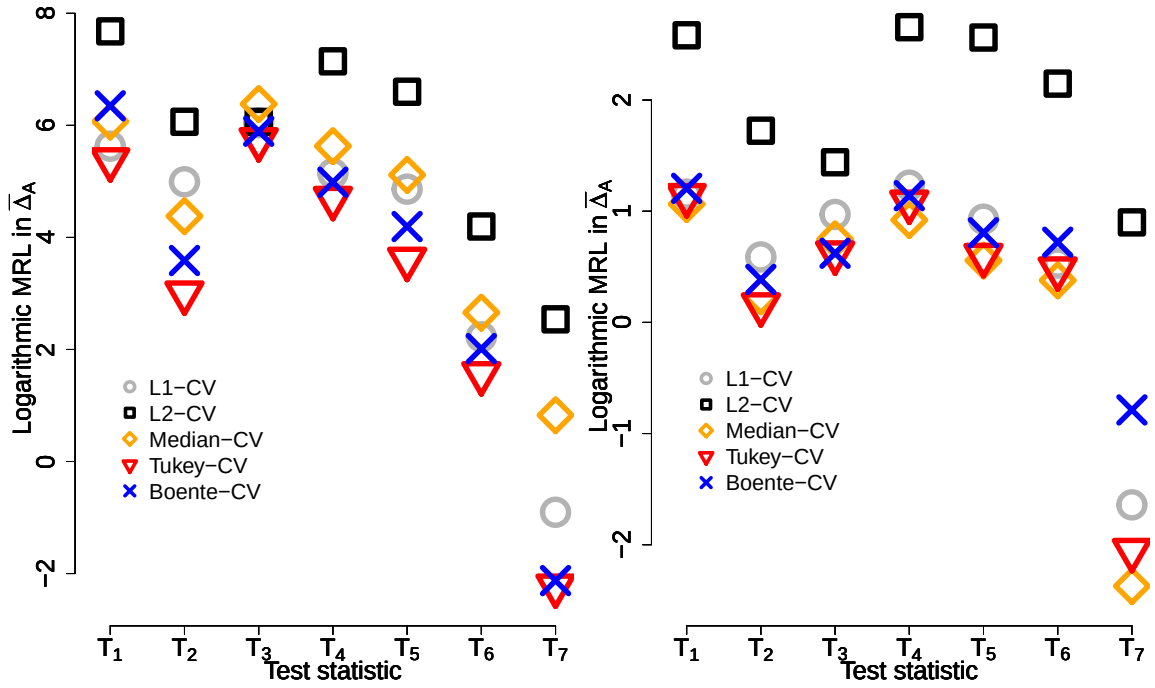

Figure 3.16: MRL in $\overline{\Delta}_A$ for the situation sets $(0.15, 192)$ (left) and $(0.30, 12)$ (right).

best, if no large outliers are observed, see Fig. 3.15 (left). Note that the t-test with Tukey-CV also delivers a small loss-value. As soon as small percentages of larger outliers are included, t-test-jump detectors become considerably worse, due to the lack of robustness of the sample mean $\tilde{\Xi}_1$ and the t-test. Detection rules based on the trimmed t-test perform best for a moderate outlier magnitude γ, see Fig. 3.15 (right). For larger γ, see Fig. 3.16 (left), or larger percentages of outliers, see Fig. 3.16 (right), the median-test delivers jump detection rules with the smallest loss-values. Tukey-CV, it is only outperformed for $\pi = 0.3$ by the highly robust median-CV-criterion. Therefore, the good performance of Tukey-CV in the direct approach can be obtained for the indirect approach. Fig. 3.16 does also show that in the presence of outliers L_2-CV performs poor, regardless of the used test-statistic.

Fig. 3.17 shows two examples, where the fitted functions of a cross-validated test statistic and a jump-preserving smoother are compared in situations with moderate sized outliers. In the left example the test misses one jump, while in the right example wrong jumps are tracked by the smoother. Applying the indirect approach gives smoother curves and so more suitable estimated function values for a piecewise constant function. However, in case of a curved f, a direct approach would be superior, due to its larger flexibility to the functional form of f. The main target of the indirect approach is a good estimation of the true jump locations and not of the true f.

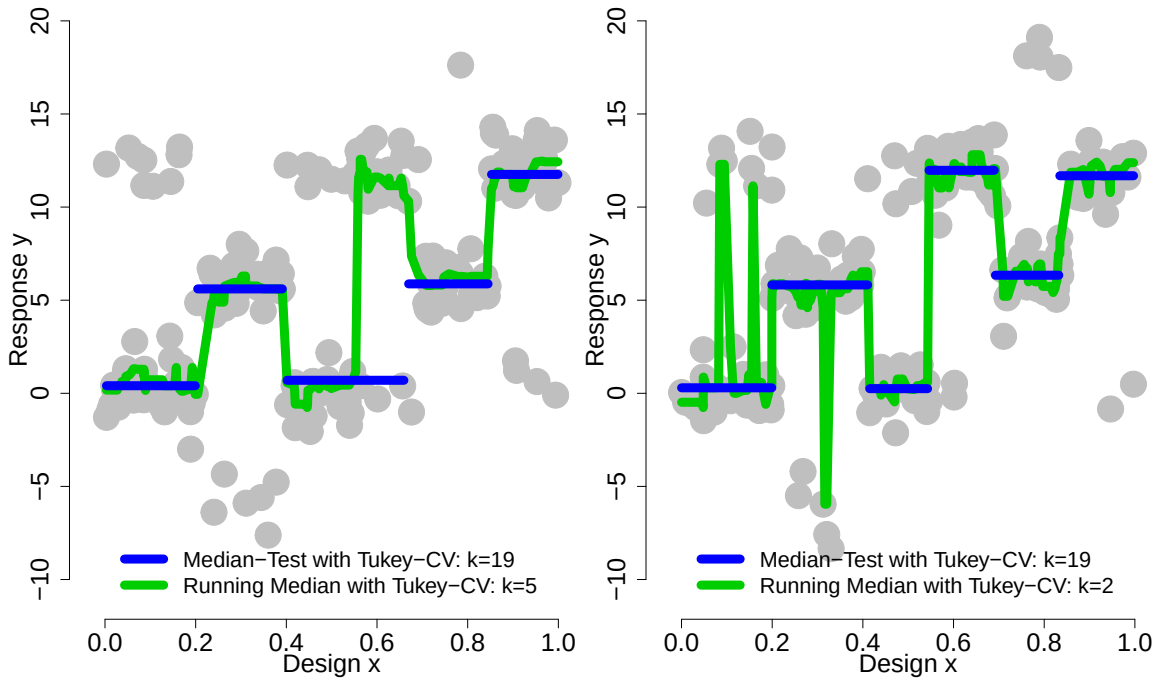

Figure 3.17: Two data examples with 15% outliers of magnitude $\gamma = 12$ and $m = 5$ jumps with height $s = 6$.

3.4 Conclusions

Outlying values are a challenge for jump detection rules based on cross-validated test statistics. The ordinary t-test and the L_2-CV perform poorly then, in terms of estimating the true jump locations and consequently the true function f. Several robust jump detection rules are compared for different data situations with outliers. The Tukey-CV-criterion delivers good results for all test statistics, whether outliers are observed or not. In the presence of large outliers the two-sample-Hodges-Lehmann-test and the median-test are appropriate choices for estimating the number of jumps and the jump locations. The latter test combined with a median estimation of the segments between the estimated jumps does also deliver good estimations of f.

To achieve further improvements, the level of significance should be chosen by the rule of Bonferroni. Furthermore, the algorithm of Wu and Chu (1993) leads to a good selection out of all jump candidates for tests based on medians and Hodges-Lehmann-estimators. For the t-tests and the rank tests the algorithm of Qiu and Yandell (1998) is preferable and an improvement to the original version is obtained, if the jump candidate with the smallest p-value of each set is taken.

Chapter 4

On nonparametric tests for trend detection in seasonal time series

4.1 Introduction

One interest in time series analysis is to detect monotonic trends in the data. Several parametric and nonparametric procedures for trend detection based on significance tests have been suggested. Parametric methods rely on strong assumptions for the distribution of the data, which are difficult to check in practice and possibly not fulfilled. Furthermore a parametric form of the trend has to be specified, where only some unknown parameters need to be estimated. Nonparametric test procedures are more flexible as they afford only rather general assumptions about the distribution. Also the trend often only needs to be monotonic without further specifications.

First ideas for nonparametric test procedures based on signs (see e.g. Cox and Stuart 1955 or Moore and Wallis 1943), ranks (see e.g. Daniels 1950 or Mann 1945) and records (Foster and Stuart, 1954) have been developed early. However, all these approaches need the assumption of i.i.d. random variables under the null hypothesis. For time series with seasonal behavior this assumption is not valid. One way to handle this problem is to estimate and subtract the seasonality. Another approach is to use tests which are robust against seasonal effects. Hirsch et al. (1982) develop a test procedure based on Kendall's test of correlation (Kendall, 1938). Diersen and Trenkler (1996) propose several tests based on records. They show that splitting the time series increases the power of the record tests, especially when seasonal effects occur. The procedures of Hirsch et. al. and Diersen and Trenkler use the independence of all observations to calculate a statistic separately for each period and sum them to get a test statistic for a test against randomness. The same ideas can be used for the above mentioned tests based on signs or ranks.

We apply the procedures to two climate time series from a gauging station in Potsdam, Germany: mean temperature and total rainfall. Such climate time series often show seasonality with a period of one year. Section 4.2 introduces the test problem of the hypothesis of randomness against a monotonic trend as well as test procedures which can also be used for seasonal data, namely some tests based on

records for the splitted time series and the seasonal Kendall–Test. We also modify other nonparametric test statistics to consider seasonality. The mentioned sign– and rank–tests are transformed to new seasonal nonparametric tests. In Section 4.3 we compare the power of the several test procedures against different types of monotone trends and in the case of autocorrelation. In Section 4.4 the two climate time series are analysed. In particular, the test procedures are used to check the hypothesis of randomness. Section 4.5 summarizes the results.

4.2 Nonparametric tests of the hypothesis of randomness $\mathcal{H}_R$

A common assumption of statistical analysis is the hypothesis of randomness. It means that some observations $x_1, \ldots, x_n$ are a realisation of independent and identically distributed (i.i.d.) continuous random variables (rvs) $X_1, \ldots, X_n$, all with the same cumulative distribution function (cdf) F. There are several test procedures which can be used to test the hypothesis of randomness H_0 against the alternative H_1 of a monotonic trend. However, in time series analysis the observations $x_1, \ldots, x_n$ are a realisation of a stochastic process and can be autocorrelated, implying a lack of independence of $X_1, \ldots, X_n$. Additionally, many time series show seasonal effects and so $X_1, \ldots, X_n$ are not identically distributed, even if there is no monotonic trend. We modify the hypothesis of randomness for seasonal data to handle at least the second problem:

Firstly, if there is a cycle of k periods, the random sample $\vec{X} = (X_1, \ldots, X_n)$ is splitted into k parts

$$\vec{X} = (\vec{X}_1, \vec{X}_2, \ldots, \vec{X}_k) \text{ with } \vec{X}_j = (X_{1,j}, X_{2,j}, \ldots, X_{n_j,j}) \text{ and } X_{i,j} = X_{k(i-1)+j} \quad (4.1)$$

for $j = 1, \ldots, k$ and $i = 1, \ldots, n_j$. $\vec{X}_j$ thus includes all n_j observations of season j. Under the null hypothesis H_0 of no trend the continuous rvs $X_1, \ldots, X_n$ are still considered to be independent but only for each j the rvs $X_{1,j}, \ldots, X_{n_j,j}$ are identically distributed with common cdf F_j. Under the alternative H_1 of a monotonic trend there are values $0 = a_{1,j} \leq a_{2,j} \leq \ldots \leq a_{n_j,j}$ with $a_{i,j} < a_{i+1,j}$ for at least one $i \in \{1, \ldots, n_j - 1\}$ and $j \in \{1, \ldots, k\}$ such that $F_{i,j}(x) = F_j(x - a_{i,j})$ in case of an increasing and $F_{i,j}(x) = F_j(x + a_{i,j})$ in case of a decreasing trend. Under H_0 the hypothesis of randomness within each period is fulfilled. In the following we denote the test problem of the hypothesis of randomness for seasonal data against a monotone trend alternative with $\mathcal{H}_R$ and introduce test procedures for $\mathcal{H}_R$.

4.2.1 Tests based on record statistics for $\mathcal{H}_R$

Foster and Stuart (1954) introduce a nonparametric test procedure for $\mathcal{H}_R$ based on the number of upper and lower records in the sequence $X_1, \ldots, X_n$ and the reversed sequence $X_n, \ldots, X_1$. A test procedure for $\mathcal{H}_R$ based on this approach which is robust

against seasonality is introduced by Diersen and Trenkler (1996). A first application of their procedure is given in Diersen and Trenkler (2001).

Using (4.1) we define upper and lower record statistics $U^o_{i,j}$, $L^o_{i,j}$, $U^r_{i,j}$ and $L^r_{i,j}$ of the original and the reversed sequence for all periods $j = 1, \ldots, k$ at $i = 2, \ldots, n_j$ as

$$U^o_{i,j} = \begin{cases} 1 & , \text{ if } X_{i,j} > \max\{X_{1,j}, X_{2,j}, \ldots, X_{i-1,j}\} \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

$$L^o_{i,j} = \begin{cases} 1 & , \text{ if } X_{i,j} < \min\{X_{1,j}, X_{2,j}, \ldots, X_{i-1,j}\} \\ 0 & \text{otherwise} \end{cases} \tag{4.3}$$

$$U^r_{n_j-i+1,j} = \begin{cases} 1 & , \text{ if } X_{n_j-i+1,j} > \max\{X_{n_j-i+2,j}, X_{n_j-i+3,j}, \ldots, X_{n_j,j}\} \\ 0 & \text{otherwise} \end{cases} \tag{4.4}$$

$$L^r_{n_j-i+1,j} = \begin{cases} 1 & , \text{ if } X_{n_j-i+1,j} < \min\{X_{n_j-i+2,j}, X_{n_j-i+3,j}, \ldots, X_{n_j,j}\} \\ 0 & \text{otherwise} \end{cases} \tag{4.5}$$

with

$$U^o_{1,j} = L^o_{1,j} = U^r_{n_j,j} = L^r_{n_j,j} = 1 \tag{4.6}$$

as the first value of a sequence is always an upper and a lower record.

Under H_0 for a larger i the probability of a record will get smaller. Therefore Diersen and Trenkler (1996) recommend to use linear weights $w_i = i - 1$ for a record at the $i-$th position of the original or reversed sequence. The sum of the weighted records of the original sequence

$$U^o = \sum_{j=1}^{k} \sum_{i=1}^{n_j} w_i U^o_{i,j} \quad \text{and} \quad L^o = \sum_{j=1}^{k} \sum_{i=1}^{n_j} w_i L^o_{i,j} \ , \tag{4.7}$$

and the sum of the records of the reversed series

$$U^r = \sum_{j=1}^{k} \sum_{i=1}^{n_j} w_i U^r_{n_j-i+1,j} \quad \text{and} \quad L^r = \sum_{j=1}^{k} \sum_{i=1}^{n_j} w_i L^r_{n_j-i+1,j} \tag{4.8}$$

can be used as test statistics for $\mathcal{H}_R$. They are sums of independent rvs and all have the same distribution under H_0. The expectations and variances are given by

$$\mathrm{E}(U^o) = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \frac{w_i}{i} \quad \text{and} \quad \mathrm{Var}(U^o) = \sum_{j=1}^{k} \sum_{i=1}^{n_j} w_i^2 \frac{i-1}{i^2} \tag{4.9}$$

and especially

$$\mathrm{E}(U^o) = k \sum_{i=1}^{n_1} \frac{i-1}{i} \quad \text{and} \quad \mathrm{Var}(U^o) = k \sum_{i=1}^{n_1} \frac{(i-1)^3}{i^2} \tag{4.10}$$

if linear weights $w_i = i - 1$ are used and all periods j have the same number of observations n_1.

If an upward trend exists, U^o and L^r become large while L^o and U^r become small. The opposite is true, if a downward trend exists. These informations can be used to combine the sums in (4.7) and (4.8) and to use the statistics

$$T_1 = U^o - L^o \ , \quad T_2 = U^o - U^r \ , \quad T_3 = U^o + L^r \ , \quad T_4 = U^o - U^r + L^r - L^o \tag{4.11}$$

for $\mathcal{H}_R$. Under H_0 the distributions of T_1, T_2 and T_3 will not change, if $\widetilde{T}_1 = L^r - U^r$, $\widetilde{T}_2 = L^r - L^o$ and $\widetilde{T}_3 = U^r + L^o$, respectively, are taken instead of the sums given in (4.11). From these statistics, only

$$T_1 = U^o - L^o \ = \ \sum_{j=1}^{k} \sum_{i=1}^{n_j} w_i \left(U^o_{i,j} - L^o_{i,j} \right) \tag{4.12}$$

can be expressed as a sum of independent rvs, because here records from the same sequence are combined. We have under H_0

$$\mathrm{E}(T_1) \ = \ 0 \quad \text{and} \quad \mathrm{Var}(T_1) \ = \ 2 \sum_{j=1}^{k} \sum_{i=1}^{n_j} \frac{w_i^2}{i} \ . \tag{4.13}$$

In contrast to T_1, in T_2, T_3 and T_4 we use records from the original sequence as well as from the reversed sequence. So the summands here are not independent. We get the expectations

$$\mathrm{E}(T_2) \ = \ \mathrm{E}(T_4) \ = \ 0 \quad \text{and} \quad \mathrm{E}(T_3) \ = \ 2 \sum_{j=1}^{k} \sum_{i=1}^{n_j} \frac{w_i}{i} \ . \tag{4.14}$$

while the variances of T_2, T_3 and T_4 become unwieldly expressions and are given in Diersen and Trenkler (2001) for the case $n_1 = \ldots = n_k$.

Diersen and Trenkler (2001) recommend a splitting with large k and small n_j, $j = 1, \ldots, k$, due to better asymptotic properties of the statistics in (4.11). With $X_1, \ldots, X_n$ assumed to be independent and $n_1 = \ldots = n_k$, the statistics T_1, T_2, T_3 and T_4 are the sum of k i.i.d. rvs. So for $k \to \infty$ all four test statistics are asymptotically normal distributed. These asymptotics are not fulfilled, if the statistics in (4.11) are only weighted but not splitted. Diersen and Trenkler (1996) showed for this case that the asymptotic distribution is not a normal one. Furthermore compared to the best parametric test in the normal linear regression model and the (non seasonal) Kendall–Test, the asymptotic relative efficiency (Noether, 1955) of the unsplitted record tests is zero, while it stays positive for an increasing splitting factor (Diersen and Trenkler, 1996, 2001). So it is also an interesting question if the performance of other nonparametric tests can be increased, if the time series is splitted with a large k and a small number n_j of observations in each period j.

4.2.2 The seasonal Kendall-test

Mann (1945) introduced a test for $\mathcal{H}_R$ based on Kendall's test for independence of two random variables in a bivariate distribution. It was modified by Hirsch et al. (1982)

to robustify the test statistic against seasonal effects. Taking the splitted series in (4.1), they use the test statistic

$$S \;=\; \sum_{j=1}^{k} S_j \qquad \text{with} \qquad S_j \;=\; \sum_{i=1}^{n_j-1} \sum_{i'=i+1}^{n_j} \operatorname{sgn}(X_{i',j} - X_{i,j}) \tag{4.15}$$

for $\mathcal{H}_R$. So in S_j the number of pairs $(X_{i,j}, X_{i',j})$ with $X_{i,j} < X_{i',j}$ is subtracted from the number of pairs $(X_{i,j}, X_{i',j})$ with $X_{i,j} > X_{i',j}$, $i < i'$, for period j. If there is a positive (negative) monotonic trend in period j, the statistic S_j is expected to be large (small) while it will probably realise a value near 0 if there is no monotonic trend. If the same positive (negative) monotonic behavior can be observed for all periods, the statistic S will also become large (small). S will also take a value close to 0, if no monotonic trend exists.

The exact distribution of S under H_0 is symmetric with

$$\mathrm{E}(S) \;=\; \sum_{j=1}^{k} \mathrm{E}(S_j) \;=\; 0 \tag{4.16}$$

and if there are no identical values (ties) in the observations of any period j, the variance is given by

$$\mathrm{Var}(S) \;=\; \sum_{j=1}^{k} \mathrm{Var}(S_j) \;=\; \sum_{j=1}^{k} \frac{n_j(n_j-1)(2n_j+5)}{18} \tag{4.17}$$

as $S_1, \ldots, S_k$ are independent. A pair of observations is called a tie of extend δ, if δ observations of $x_1, \ldots, x_n$ have the same value. If $X_1, \ldots, X_n$ are continuous rvs, the probability of a tie is zero, but for rounded values, ties can be observed. Let $n_{\delta,j}$ be the number of ties within $\vec{X}_j$ with extend δ. Then the variance of S becomes smaller:

$$\mathrm{Var}(S) \;=\; \sum_{j=1}^{k} \frac{\left(n_j(n_j-1)(2n_j+5) - \sum_{\delta=1}^{n_j} n_{\delta,j}\delta(\delta-1)(2\delta+5) \right)}{18} \tag{4.18}$$

As every S_j is asymptotically normally distributed for $n_j \to \infty$, the statistic S as a finite sum of independent asymptotically normally distributed rvs is asymptotically normal, too, if n_j converges to infinity for each j. The exact distribution of S under H_0 (neglecting ties) can be determined by enumerating all permutations of $X_{1,j}, \ldots, X_{n_j,j}$ for each j and calculating the values of S_j for every permutation of each j. The individual values and their frequencies can be easily calculated with Chapter 5 of Kendall and Gibbons (1990). According to the frequencies of the single values for each S_j, the distribution of S can be obtained by reconsidering every possible combination of the values and multiplying the corresponding frequencies. However, for large n calculating the exact distribution of S is time consuming, so the normal approximation should be used whenever possible. Hirsch et al. (1982) state that already for $k = 12$

and $n_j = 3$ the normal approximation of S_j works well. They also claim that their test is robust against seasonality and departures from normality, but not robust against dependence. Hirsch and Slack (1984) develop a test for $\mathcal{H}_R$, which performs better than S if the data are autocorrelated. This test uses estimates of the covariances between two seasons based on Spearman's rank correlation coefficient. The estimated covariances are used to correct the variance of S in the normal approximation.

4.2.3 Tests based on rank statistics for $\mathcal{H}_R$

Aiyar et al. (1979) compare the asymptotic relative efficiencies of many nonparametric tests for the hypothesis of randomness against trend alternatives. They consider mostly linear and nonlinear rank statistics, which we will use in the following for $\mathcal{H}_R$:

Taken the splitted series from (4.1) let $R(X_{1,j}), \ldots, R(X_{n_j,j})$ be the ranks of the continuous rvs $X_{1,j}, \ldots, X_{n_j,j}$, for $j \in \{1, \ldots, k\}$. Then two linear rank test statistics based on Spearman's rank correlation coefficient are given by

$$R_1 = \sum_{j=1}^{k} \widetilde{R}_{1,j} \text{ with } \widetilde{R}_{1,j} = \sum_{i=1}^{n_j} \left(i - \frac{n_j+1}{2} \right) \left(R(X_{i,j}) - \frac{n_j+1}{2} \right) \tag{4.19}$$

and

$$R_2 = \sum_{j=1}^{k} \widetilde{R}_{2,j} \text{ with } \widetilde{R}_{2,j} = \sum_{i=1}^{n_j} \left(i - \frac{n_j+1}{2} \right) \text{sign} \left(R(X_{i,j}) - \frac{n_j+1}{2} \right) . \tag{4.20}$$

Both statistics are symmetric and have an expected value of 0. Their variances are

$$\text{Var}(R_1) = \sum_{j=1}^{k} \text{Var}(\widetilde{R}_{1,j}) = \sum_{j=1}^{k} \frac{n_j^2 (n_j+1)^2 (n_j-1)}{144} \tag{4.21}$$

and

$$\begin{aligned} \text{Var}(R_2) &= \sum_{j=1}^{k} \text{Var}(\widetilde{R}_{2,j}) \text{ with} \qquad (4.22) \\ \text{Var}(\widetilde{R}_{2,j}) &= \begin{cases} \displaystyle\sum_{j=1}^{k} \frac{n_j^2 (n_j+1)}{12} & , \ n_j \text{ even} \\ \displaystyle\sum_{j=1}^{k} \frac{n_j (n_j-1)(n_j+1)}{12} & , \ n_j \text{ odd} . \end{cases} \end{aligned}$$

Instead of considering all rvs like in (4.19) and (4.20), the $(1-2\gamma)$ truncated sample can be taken for all periods, with $\gamma \in (0, 0.5)$. Like Aiyar et al. (1979) we define

$$c_{i,j} = \begin{cases} -1 & , \quad 0 < i \leq \lfloor \gamma n_j \rfloor \\ 0 & , \quad \lfloor \gamma n_j \rfloor < i \leq n_j - \lfloor \gamma n_j \rfloor \\ +1 & , \quad n_j - \lfloor \gamma n_j \rfloor < i \leq n_j \end{cases} \tag{4.23}$$

so that the two statistics

$$R_3 = \sum_{j=1}^{k} \widetilde{R}_{3,j} \quad \text{with} \tag{4.24}$$

$$\widetilde{R}_{3,j} = \sum_{i=1}^{n_j} c_{i,j} \left(R(X_{i,j}) - \frac{n_j+1}{2} \right) = \sum_{j=1}^{k} \left(\sum_{i=n_j-\lfloor \gamma n_j \rfloor+1}^{n_j} R(X_{i,j}) - \sum_{i=1}^{\lfloor \gamma n_j \rfloor} R(X_{i,j}) \right)$$

and

$$R_4 = \sum_{j=1}^{k} \widetilde{R}_{4,j} \quad \text{with} \tag{4.25}$$

$$\begin{aligned} \widetilde{R}_{4,j} &= \sum_{i=1}^{n_j} c_{i,j} \operatorname{sign} \left(R(X_{i,j}) - \frac{n_j+1}{2} \right) \\ &= \sum_{i=n_j-\lfloor \gamma n_j \rfloor+1}^{n_j} \operatorname{sign} \left(R(X_{i,j}) - \frac{n_j+1}{2} \right) - \sum_{i=1}^{\lfloor \gamma n_j \rfloor} \operatorname{sign} \left(R(X_{i,j}) - \frac{n_j+1}{2} \right) \end{aligned}$$

compare the sum of the most recent $\lfloor \gamma n_j \rfloor$ ranks (signs) with the sum of the first $\lfloor \gamma n_j \rfloor$ ranks (signs). Again the expectation of R_3 and R_4 is 0. Under the null hypothesis, the variances are given by

$$\operatorname{Var}(R_3) = \sum_{j=1}^{k} \frac{n_j(n_j+1)\lfloor \gamma n_j \rfloor}{6} \quad \text{and} \tag{4.26}$$

$$\operatorname{Var}(R_4) = \sum_{j=1}^{k} \operatorname{Var}(\widetilde{R}_{4,j}) \quad \text{with} \tag{4.27}$$

$$\operatorname{Var}(\widetilde{R}_{4,j}) = \begin{cases} 2\dfrac{n_j}{n_j-1}\lfloor \gamma n_j \rfloor & , \quad n_j \text{ even} \\ 2\lfloor \gamma n_j \rfloor & , \quad n_j \text{ odd} \end{cases} .$$

Again the above variances are only valid if all observations have different values. If ties occur, one possibility, which leads to a loss of power but keeps the variances from (4.23) and (4.28) under the null hypothesis is to give random ranks to tied observations. Alternatives like average ranks, which reduce the loss of power compared to random ranks, are not considered here.

In addition to this, Aiyar et al. (1979) also consider nonlinear rank statistics. In analogy to them we define for each period j

$$\mathbb{1}_{i,i',j} = \begin{cases} 1 & , \quad \text{if} \quad X_{i,j} < X_{i',j} \\ 0 & , \quad \text{otherwise} \end{cases} , \tag{4.28}$$

$i, i' \in \{1, \ldots, n\}, i \neq i'$. Under the null hypothesis of randomness, we have

$$\mathrm{E}(\mathbb{1}_{i,i',j}) = \frac{1}{2} \quad \text{and} \quad \operatorname{Var}(\mathbb{1}_{i,i',j}) = \frac{1}{4} . \tag{4.29}$$

Based on the sign difference test Moore and Wallis (1943) we define for $\mathcal{H}_R$

$$N_1 = \sum_{j=1}^{k} \widetilde{N}_{1,j} \quad \text{with} \quad \widetilde{N}_{1,j} = \sum_{i=2}^{n_j} \mathbb{1}_{i-1,i,j} \tag{4.30}$$

which counts the number of pairs for each period j, where the consecutive observation has a larger value and then sums these pairs over all periods. For each j we have $n_j - 1$ differences. Under H_0 and from (4.29) we get

$$\mathrm{E}(N_1) = \sum_{j=1}^{k} \frac{1}{2}(n_j - 1) \quad \text{and} \quad \mathrm{Var}(N_1) = \sum_{j=1}^{k} \frac{1}{12}(n_j + 1) \ . \tag{4.31}$$

For each j the distribution of $\widetilde{N}_{1,j}$ converges to a normal distribution (Moore and Wallis, 1943). Therefore N_1 is asymptotically normally distributed, too.

Another test for $\mathcal{H}_R$ based on Cox and Stuart (1955) is given by

$$N_2 = \sum_{j=1}^{k} \widetilde{N}_{2,j} \quad \text{with} \quad \widetilde{N}_{2,j} = \sum_{i=1}^{\lfloor n_j/2 \rfloor} (n_j - 2i + 1)\mathbb{1}_{i,n_j-i+1,j} \ . \tag{4.32}$$

Cox and Stuart (1955) show that N_2 leads to the best weighted sign test with respect to the efficiency of a sign test of $\mathcal{H}_R$. The linear rank test statistics R_1 and R_2 and the procedure S of Kendall compare all pairs of observations, while in (4.32) each observation is taken only for one comparison. Using (4.29) we get under H_0

$$\begin{aligned} \mathrm{E}(N_2) = & \sum_{j=1}^{k} \mathrm{E}(\widetilde{N}_{2,j}) \qquad \text{with} \ \ \mathrm{E}(\widetilde{N}_{2,j}) = \begin{cases} \frac{n_j^2}{8} & , \ n_j \text{ even} \\ \frac{n_j^2-1}{8} & , \ n_j \text{ odd} \end{cases} \\ \text{and} \ \ \mathrm{Var}(N_2) = & \sum_{j=1}^{k} \tfrac{1}{24} n_j(n_j^2 - 1) \ . \end{aligned} \tag{4.33}$$

Cox and Stuart (1955) also introduce a best unweighted sign test, which can be formulated for $\mathcal{H}_R$ as follows

$$N_3 = \sum_{j=1}^{k} \widetilde{N}_{3,j} \quad \text{with} \quad \widetilde{N}_{3,j} = \sum_{i=1}^{\nu_j} \mathbb{1}_{i,n_j-\nu_j+i,j} \ . \tag{4.34}$$

The value $\nu_j \leq \frac{1}{2} n_j$ is taken to compare observations further apart. We get

$$\mathrm{E}(N_3) = \sum_{j=1}^{k} \frac{\nu_j}{2} \quad \text{and} \quad \mathrm{Var}(N_3) = \sum_{j=1}^{k} \frac{\nu_j}{4} \tag{4.35}$$

under H_0. Cox and Stuart (1955) recommend $\nu_j = \frac{1}{3} n_j$.

Again a splitting with small $n_1 = \ldots = n_k$ and large k leeds asymptotically to a normal distribution for all introduced test statistics, as k i.i.d. rvs are added.

4.3 Comparison of the nonparametric tests

4.3.1 Robustness against seasonality

Now we compare the different tests presented in Section 4.2 for different sample sizes and splitting factors and for various alternatives. We consider the time series model

$$X_{i,j} = a_{i,j} + E_{i,j} \quad j = 1, \ldots, k, \; i = 1, \ldots, n_j, \tag{4.36}$$

where $E_{1,1}, \ldots, E_{n_k,k}$ are Gaussian white noise with expected value 0 and constant variance $\sigma_E^2 = 1$. $X_{i,j}$ is the $i-$th observation for season j. For simplicity we fix the number of seasons to $k = 4$ and assume that each season has the same sample size n_1. Furthermore, the slopes are given by $a_{1,j} \leq \ldots \leq a_{n_1,j}$. We are interested in particular in three different kinds of monotone trends, with the same trend structure in each season. This means that for each j we have the same slopes. With $a_{i,j} = i\theta$ we achieve a linear trend, where the parameter θ controls the slope of the straight line. We also consider a concave case with $a_{i,j} = \theta\sqrt{n_1 i}$, and a convex case with $a_{i,j} = \theta i^2/n_1$, so that all trends increase to θn_1. We consider sample sizes $n \in \{12, 24, 32, 48, 64, 96, 120\}$ and splittings into $\widetilde{k} \in \{1, 4, 8, 12, 16, 24, 32\}$ groups whenever $\widetilde{n}_1 = n/\widetilde{k}$ is an integer. We do not consider splittings with $\widetilde{n}_1 = 2$ as here R_3 and R_4 for $\gamma = \frac{1}{3}$ as well as N_3 with $\nu_1 = \ldots = \nu_{\widetilde{k}} = \frac{1}{3}$ are not defined. The other test statistics are equivalent in this case, as they all consider an unweighted ranking of two observations in each splitting. With $\widetilde{k} = 1$ the unsplitted case is also

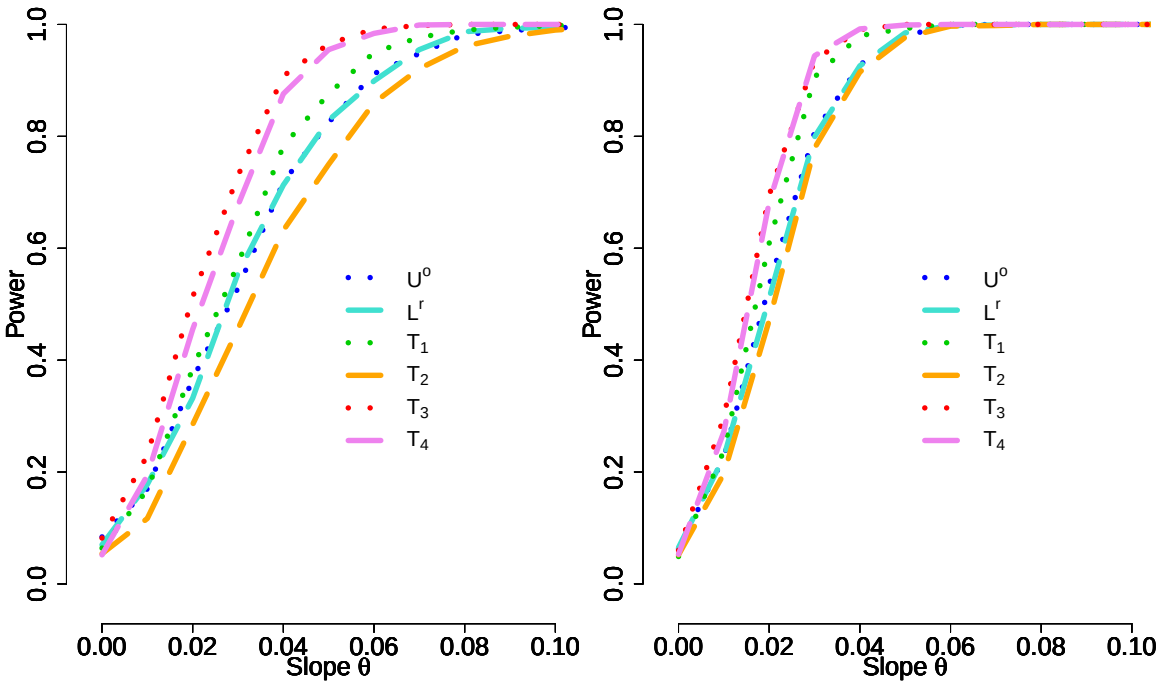

Figure 4.1: Power functions of the record tests for $n = 64$, small θ and $\widetilde{k} = 1$ (left) and $\widetilde{k} = 4$ (right) for linear trends.

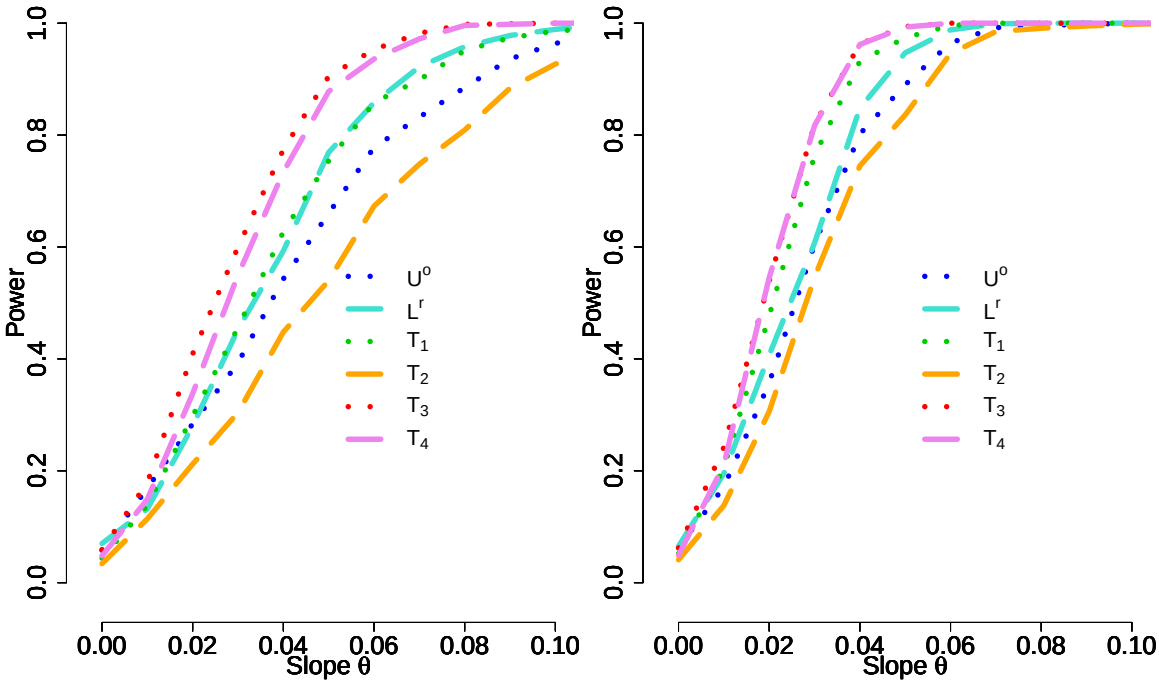

Figure 4.2: Power functions of the record tests for $n = 64$, small θ and $\widetilde{k} = 1$ (left) and $\widetilde{k} = 4$ (right) for concave trends.

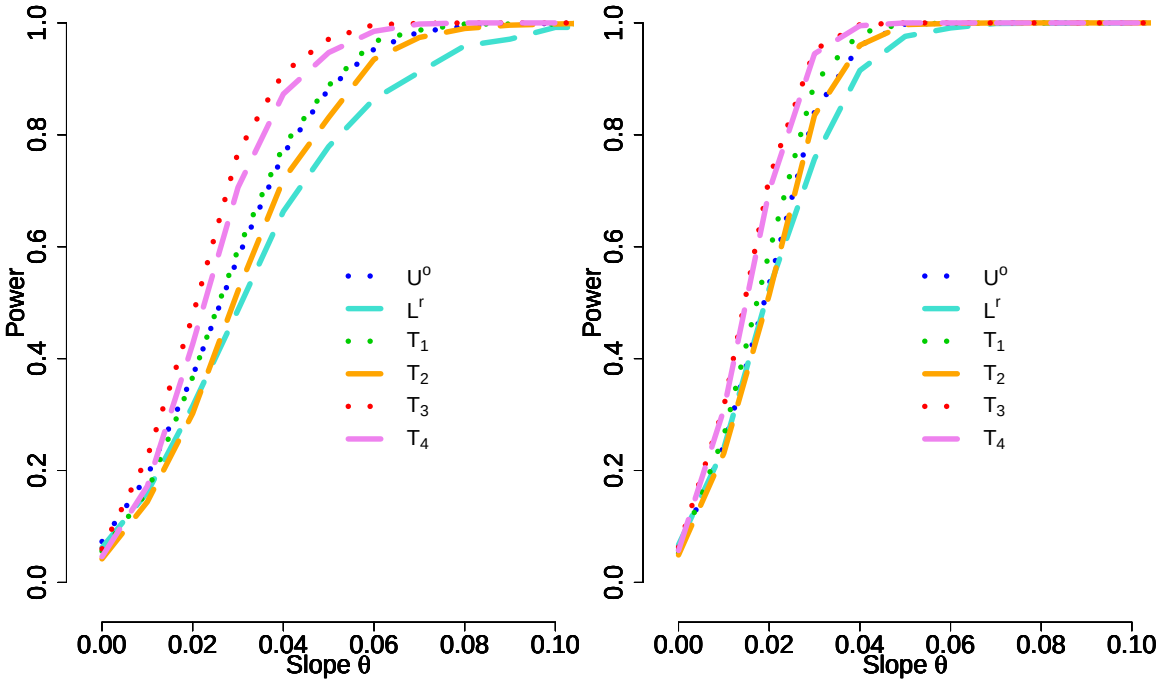

Figure 4.3: Power functions of the record tests for $n = 64$, small θ and $\widetilde{k} = 1$ (left) and $\widetilde{k} = 4$ (right) for convex trends.

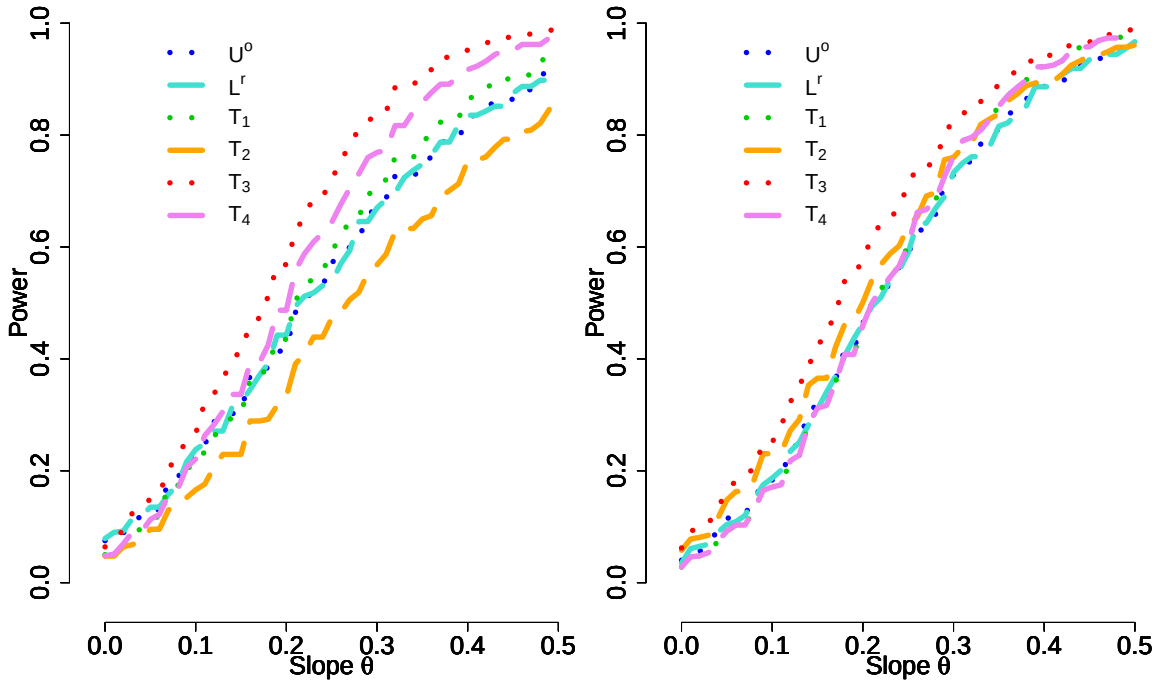

Figure 4.4: Power functions of the record tests for $n = 12$ (top) with $\widetilde{k} = 1$ (left) and $\widetilde{k} = 4$ (right).

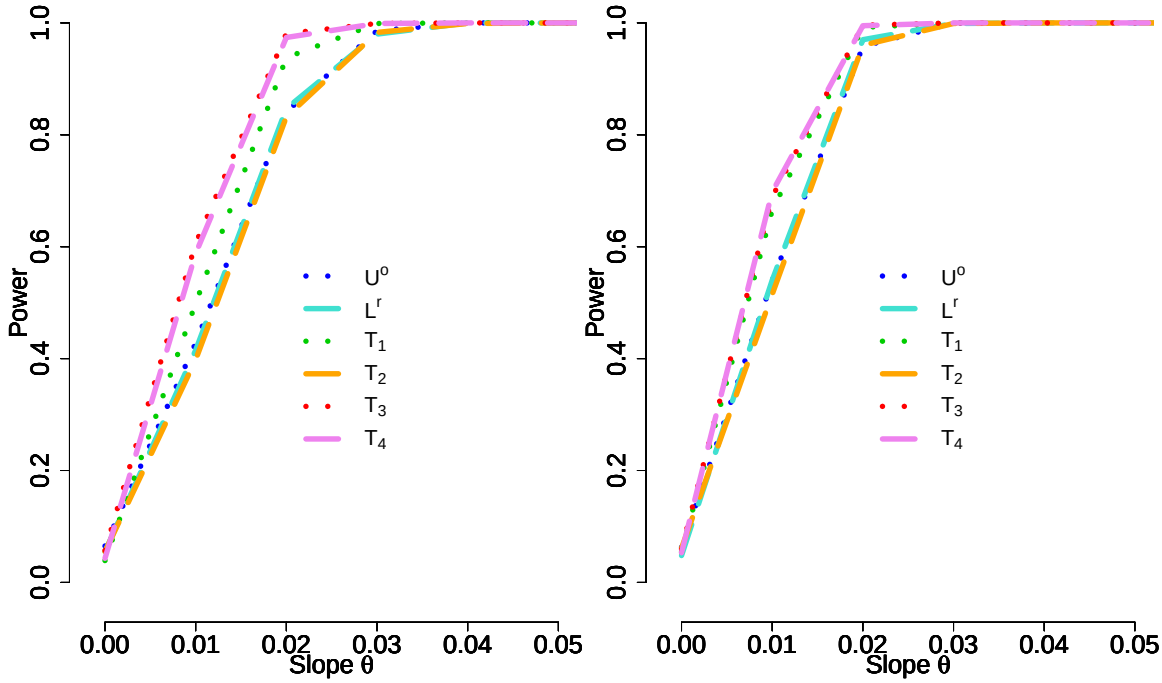

Figure 4.5: Power functions of the record tests for $n = 96$ (top) with $\widetilde{k} = 4$ (left) and $\widetilde{k} = 12$ (right).

taken into account. In case of seasonal effects the power of all tests will probably be reduced if $\widetilde{k} = 1$ is chosen. We compare the power of the tests of Section 4.2 for all reasonable combinations of $\widetilde{k}$ and n from above and take 1000 random samples from (4.36) for each combination. The percentage cases of rejections of H_0 estimate the power of the several test procedures. Here we only consider the case of an upward trend, i.e. $\theta > 0$.

We consider the linear, the convex and the concave case from above and calculate the power of all tests for $\theta \in \{0.01, 0.02, \ldots, 0.49, 0.50\}$. To achieve monotone power functions, we use the `R`–function `isotone` from the `R`–package `EbayesThresh` for monotone least squares regression to smooth the simulated power curves (R Development Core Team, 2011; Silverman, 2005).

Firstly we compare the weighted record statistics. For $n \gtrsim 64$ all power functions take values close to 1, independently of the splitting factor $\widetilde{k}$, if a linear trend with $\theta > 0.1$ exists. In the concave case only U^o and T_2 with $\widetilde{k} = 1$ perform worse for $n = 64$. An explanation for this is the strength of the slope. A positive concave trend increases less towards the end of the time series. Hence there will be fewer records at the end of the time series and U^o will perform worse than L^r. As our version of T_2 also uses U^o we receive similar results for this test statistic. In the convex case similar results can be obtained for L^r as a convex upward trend of the original sequence means a concave downward trend of the negative reversed series. The power functions of the record tests for $\widetilde{k} = 1$ and $\widetilde{k} = 4$ can be seen in Fig. 4.1 for the linear, in Fig. 4.2 for the concave and in Fig. 4.3 for the convex case. Looking also at other sample sizes n in the linear case (see Fig. 4.4 and 4.5), we find that T_3 performs best

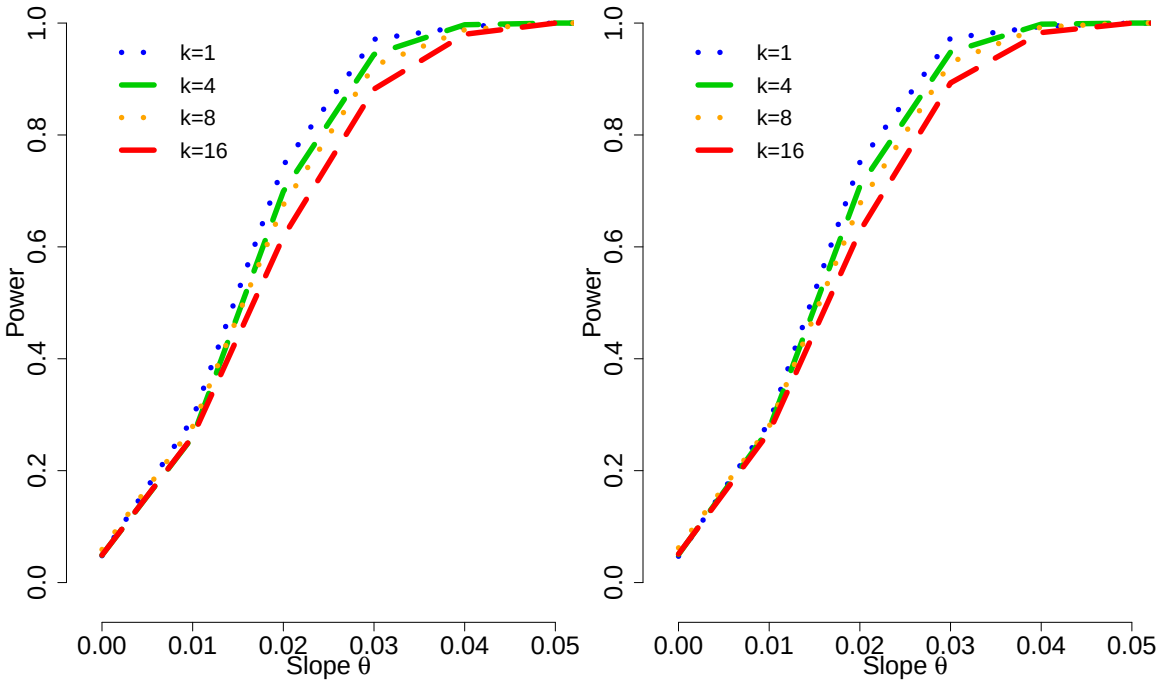

Figure 4.6: Power functions for S (left) and R_1 (right) for different $\widetilde{k}$ with $n = 64$ and a concave trend.

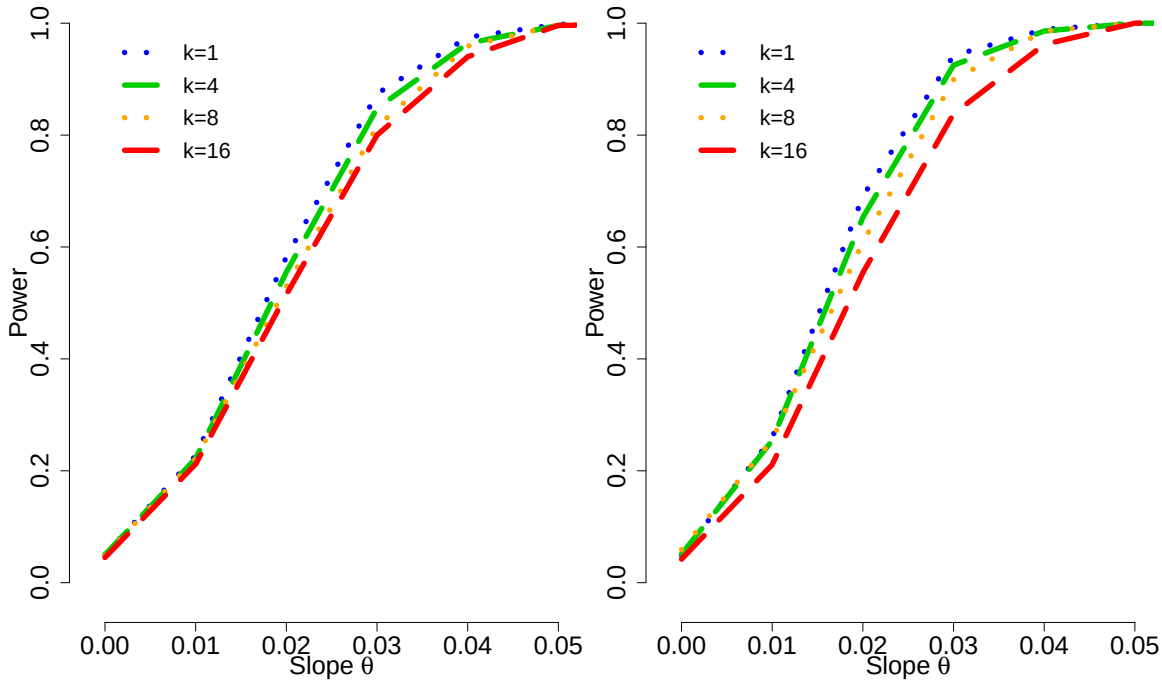

Figure 4.7: Power functions for R_2 (left) and R_3 (right) for different $\widetilde{k}$ with $n = 64$ and a concave trend.

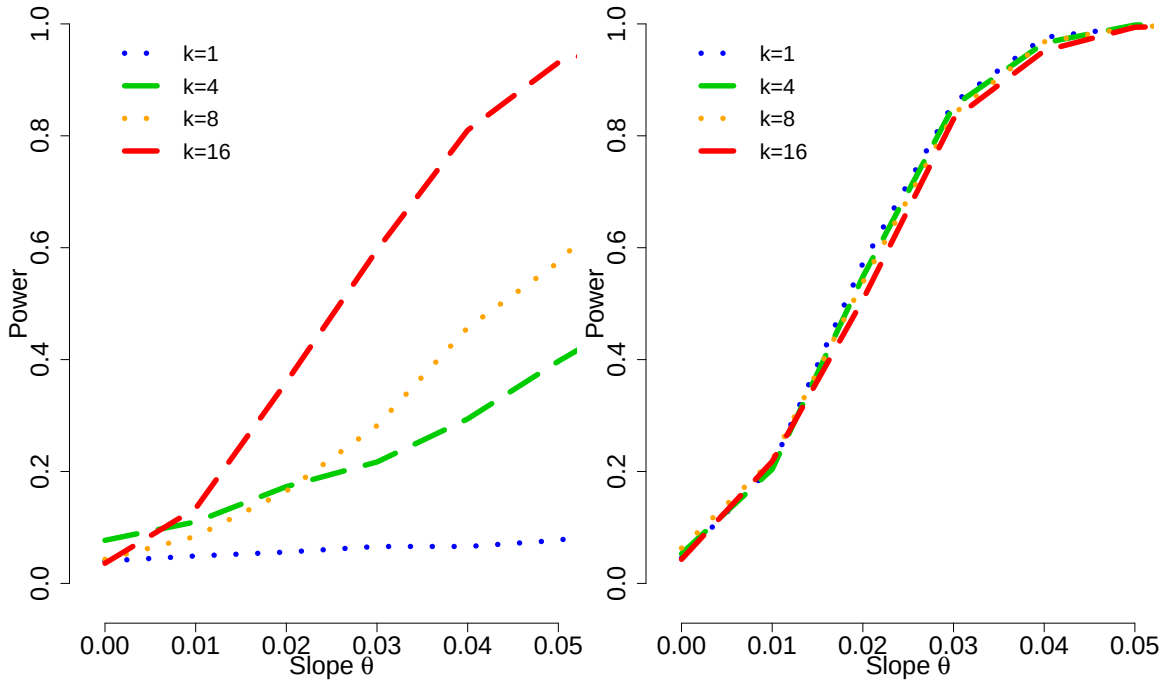

Figure 4.8: Power functions for N_1 (left) and N_2 (right) for different $\widetilde{k}$ with $n = 64$ and a concave trend.

among the record tests in most of the cases. Generally, the power of the record tests gets larger in the above situations, if a larger $\widetilde{k}$ is chosen. Only T_3 performs better for a medium value of $\widetilde{k}$, e.g. $\widetilde{k} = 4$ for $n = 32$ or $\widetilde{k} = 12$ for $n = 96$. The previous findings are confirmed in the case of a convex or concave trend.

In Fig. 4.6, 4.7 and 4.8 the power functions of the rank tests are shown, when different $\widetilde{k}$ for a fixed $n = 64$ are used. We show the concave case here, because the differences are qualitatively the same, but slightly bigger than for the linear or the convex trend. The seasonal Kendall–Test S and Spearman–Test R_1 perform best, when a small $\widetilde{k}$ is used. Conclusions about an optimal splitting for the other rank tests are hard to state. If $\widetilde{k}$ is large compared to n, the power of the tests is reduced for most of the situations. However, generally we observe for all these tests (except N_1) good results, if $\widetilde{k} = 4$ is chosen. N_1 performs worse than the other tests in most situations even though it is the only test statistic with an increasing power in case of a larger splitting factor $\widetilde{k}$. From the rank tests S and R_1 achieve the largest power in most situations. Comparing the best rank tests S and R_1 with $\widetilde{k} = 4$ and the best record tests T_3 and T_4 with a large splitting factor $\widetilde{k} = 4$, S and R_1 have a larger power in every situation.

4.3.2 Robustness against autocorrelation

Next we consider a situation with autocorrelated data. Here the hypothesis of randomness is not fulfilled, but no monotone trend exists. It is interesting which test procedures are sensitive to autocorrelation in the sense that they reject H_0 even though there is no monotone trend. We consider an autoregressive process of first order (AR(1))

$$E_t = \varrho E_{t-1} + \varepsilon_t \ , \quad t = 1, \ldots, n \ , \tag{4.37}$$

with autocorrelation coefficient ϱ, i.e. we assume the sequence $E_1, \ldots, E_n$ to be autocorrelated with correlation ϱ and hence the autocorrelation within $E_{1,j}, \ldots, E_{n_1,j}$ with $E_{i,j} = E_{k(i-1)+j}$ is smaller than ϱ. The innovations $\varepsilon_{1,j}, \ldots, \varepsilon_{n_1,j}$ are i.i.d. normally distributed random variables with expectation 0 and variance σ_ε^2, where

$$\sigma_\epsilon^2 = (1 - \varrho^2)\sigma_E^2 = (1 - \varrho^2) \tag{4.38}$$

as we want to keep σ_E^2 equal to 1 again. We vary ϱ in $\{0.025, 0.05, \ldots, 0.875, 0.9\}$.

The resulting detection rates of the record tests can be seen in Fig. 4.9 and 4.10 for $n = 96$ and different values of $\widetilde{k}$. T_3 is more sensitive to positive autocorrelation than T_1, T_2 and T_4 if a small $\widetilde{k}$ is used, but this difference vanishes for a large $\widetilde{k}$. The better performance of T_1, T_2 and T_4 for small $\widetilde{k}$ can be explained by the fact that they subtract statistics which become large in case of monotonically decreasing sequences from statistics which become large in case of monotonically increasing sequences. Positive autocorrelations cause both patterns to occur so that the effects cancel out.

For the rank tests we get the following findings, compare Fig. 4.11 and 4.12: N_2 seem to be less sensitive against autocorrelations $\varrho \leq 0.6$ for larger sample sizes $n \geq 48$, if we choose $\widetilde{k}$ so that we have three observations in each split. We observe for the pairs $n = 48$, $\widetilde{k} = 16$ and $n = 96$, $\widetilde{k} = 32$ for most of the values of ϱ a detection

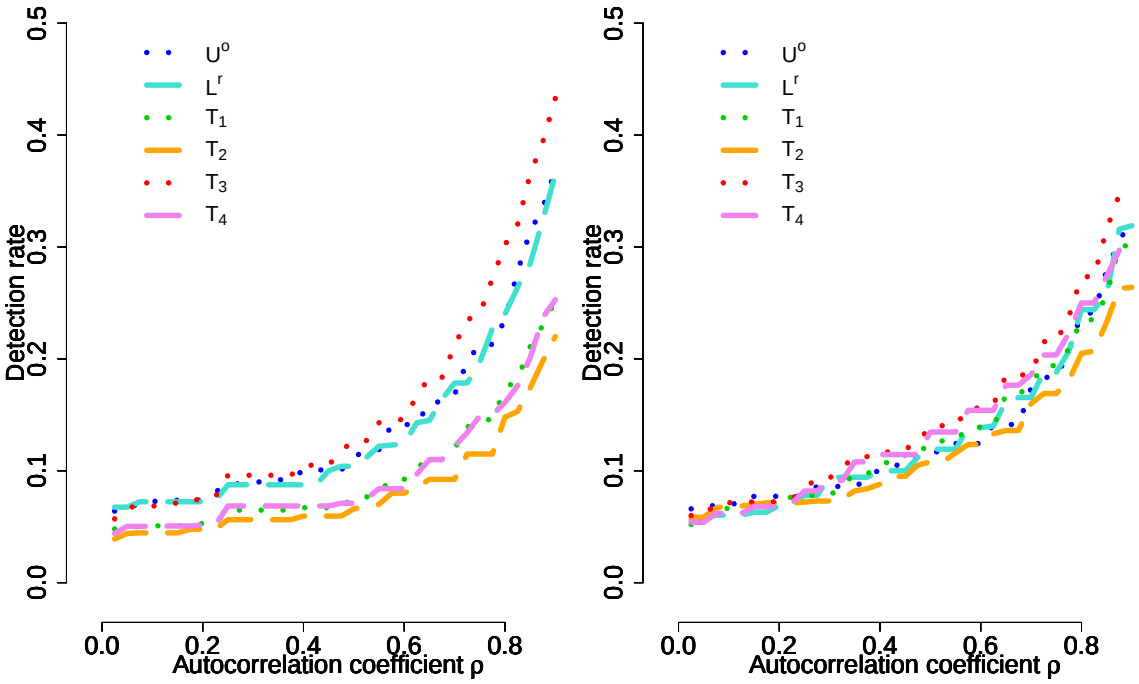

Figure 4.9: Detection rates of the record tests for $n = 96$ with $\widetilde{k} = 1$ (left) and $\widetilde{k} = 4$ (right) for autocorrelated series.

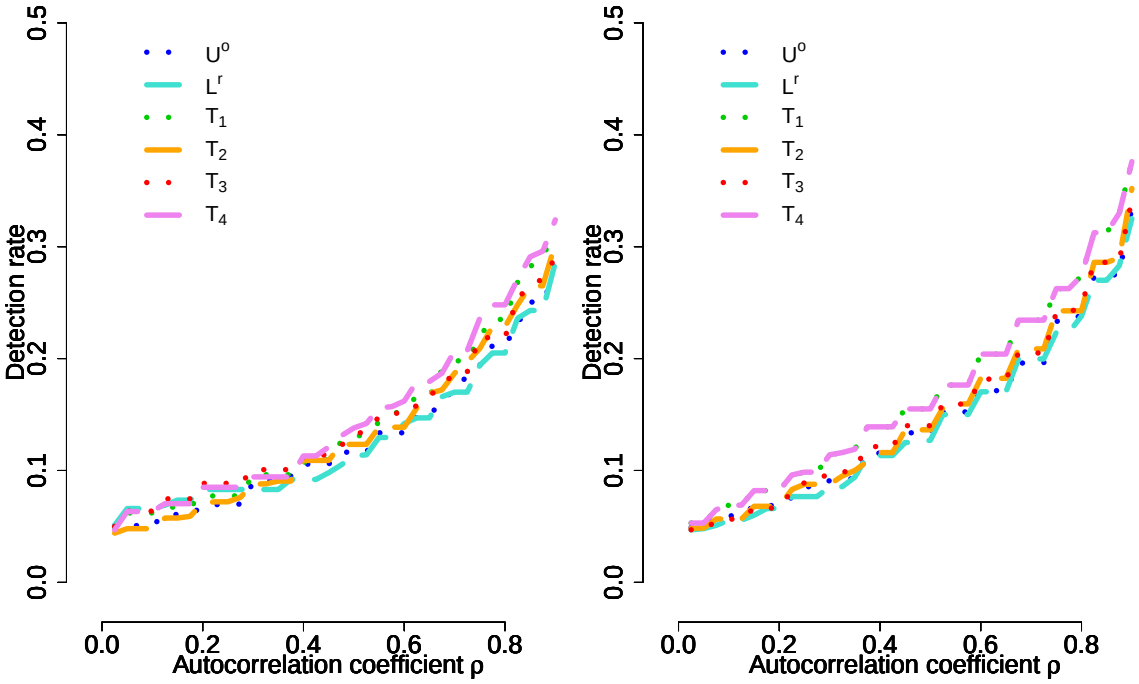

Figure 4.10: Detection rates of the record tests for $n = 96$ with $\widetilde{k} = 16$ (left) and $\widetilde{k} = 32$ (right) for autocorrelated series.

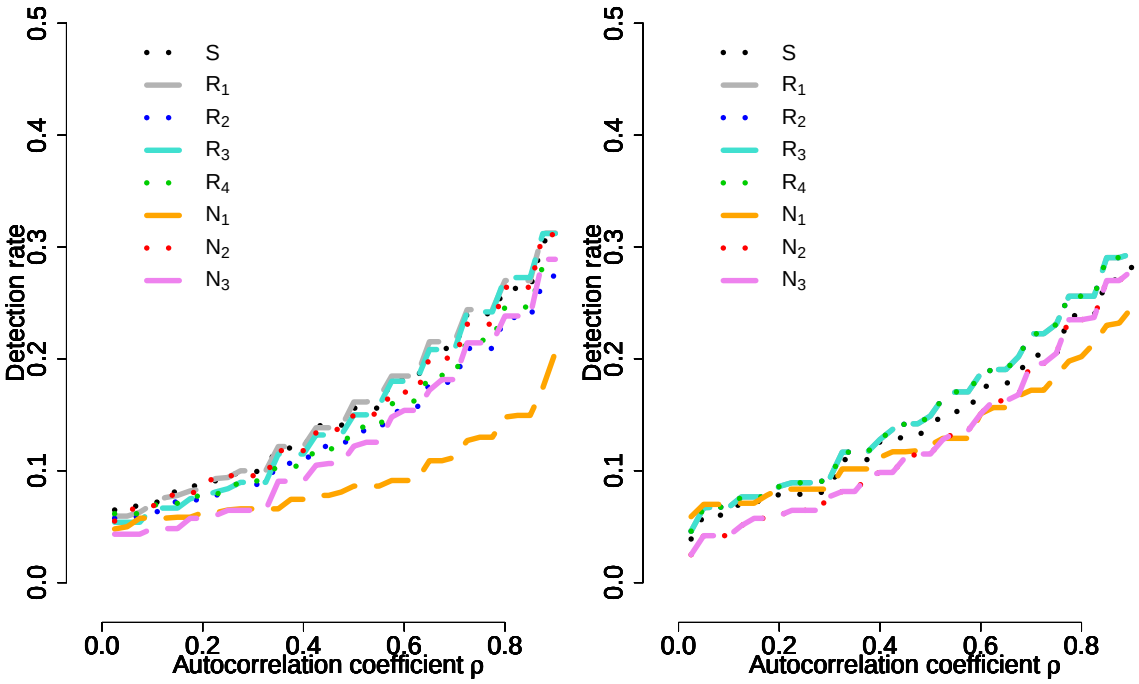

Figure 4.11: Detection rates of the rank tests for $n = 24$ with $\widetilde{k} = 4$ (left) and $\widetilde{k} = 8$ (right) with autocorrelation.

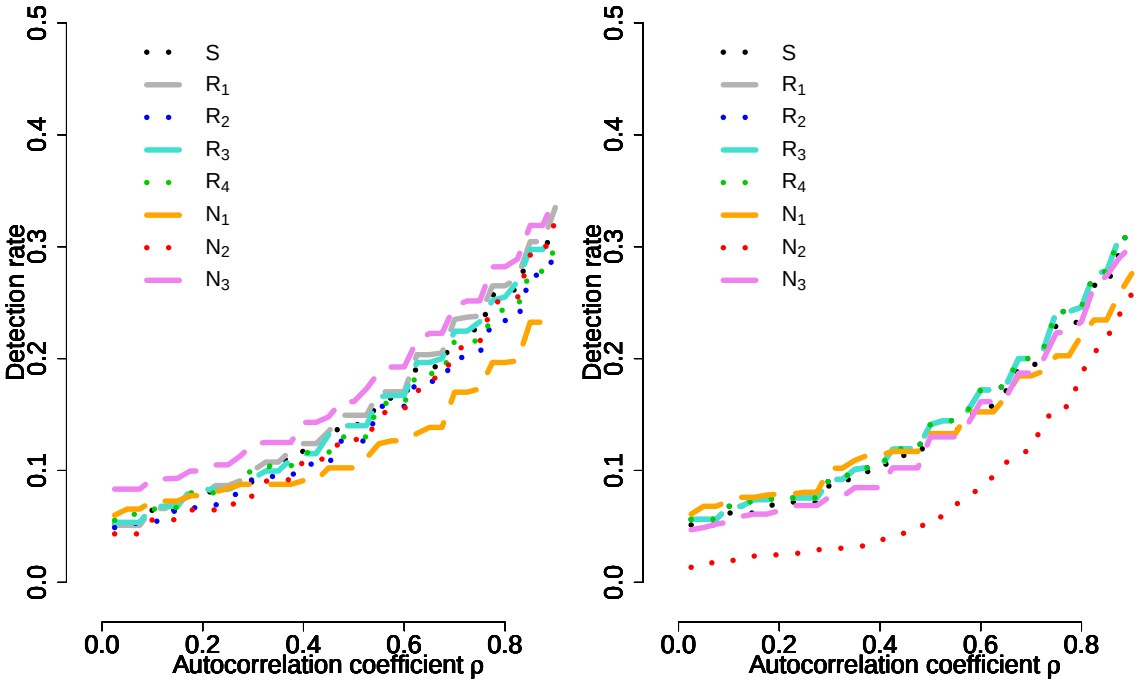

Figure 4.12: Detection rates of the rank tests for $n = 48$ with $\widetilde{k} = 12$ (left) and $\widetilde{k} = 16$ (right) with autocorrelation.

rate of less than $\alpha = 0.05$. If we choose a splitting factor leading to $n_1 > 3$ the better performance of N_2 is lost. N_1 behaves the most insensitive against autocorrelation for a large $\widetilde{k}$, but N_1 was also the test with the smallest power if a trend exists. For the other tests we have for a fixed n a higher detection rate, when a smaller splitting factor $\widetilde{k}$ is used. If we compare the record tests with the rank tests, we find that the record tests react less sensitive to autocorrelation than the rank tests in most situations.

4.4 Analysis of the climate time series from Potsdam

Now the methods from Section 4.2 are applied to some real time series data. The two series analysed here consist of the monthly observations of the mean air temperature and the total rainfall in Potsdam between January 1893 and April 2008. There are no missing values. The secular station in Potsdam is the only meteorological station in Germany for which daily data have been collected during a period of over 100 years. The measures are homogeneous, what is due to the facts that the station has never changed its position, the measuring field stayed identical and the sort of methods, prescriptions and instruments, which are used for the measuring, have been kept.

Before the methods from Section 4.2 can be applied, we have to check if the assumptions are fulfilled. Independence of the observations can be checked with the autocorrelation function (ACF) and the partial autocorrelation function (PACF). Before this we detrend the time series by subtracting a linear trend. We also deseasonalize the time series by estimating and subtracting a seasonal effect for each month. The original and the detrended deseasonalized time series can be found in Fig. 4.13 and 4.14, respectively, for the total rainfall and in Fig. 4.15 and 4.16 for the mean temperature, respectively. The autocorrelation functions of the detrended and deseasonalized time series do not show correlation in case of the rainfall (see Fig. 4.17) and positive autocorrelations at small time lags in case of the temperature (see Fig. 4.18). For the temperature series, a first order autoregressive model with a moderately large AR(1) coefficient gives a possible description of the correlations. We use the test statistics from Section 4.2 to test the hypothesis of randomness against the alternative of an upward trend in both time series.

We consider all test statistics except L^o and U^r as these tests are only useful to detect a downward trend. As we have in both time series monthly observations for more than 115 years, we choose the splitting factor $\widetilde{k}$ as multiples of 12, more precisely $\widetilde{k} \in \{12, 24, 60, 120, 240, 360\}$. This guarantees that even R_3, R_4 (with $\gamma = \frac{1}{3}$) and N_3 (with $\nu_j = \frac{1}{3}n_j$) can be computed for each split. For every test procedure we use the asymptotic critical values, which seems to be reasonable for the above $\widetilde{k}$. The resulting p–values can be seen in Table 4.1 for the total rainfall time series and in Table 4.2 for the mean temperature.

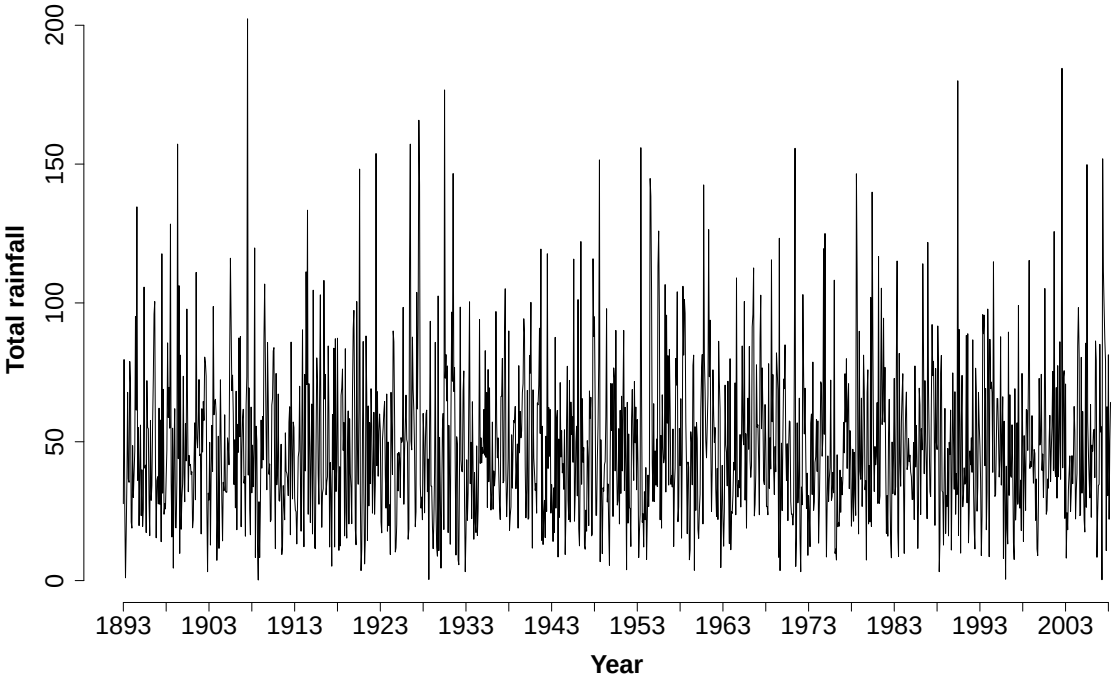

Figure 4.13: Original total rainfall time series.

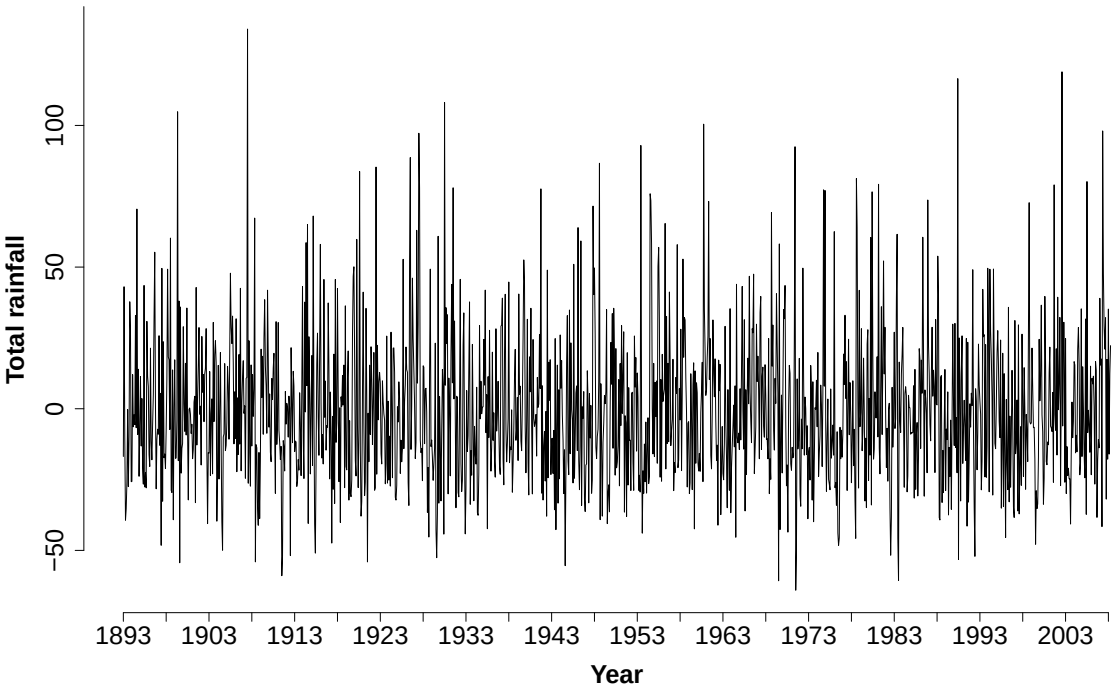

Figure 4.14: Detrended and deseasonalized total rainfall time series.

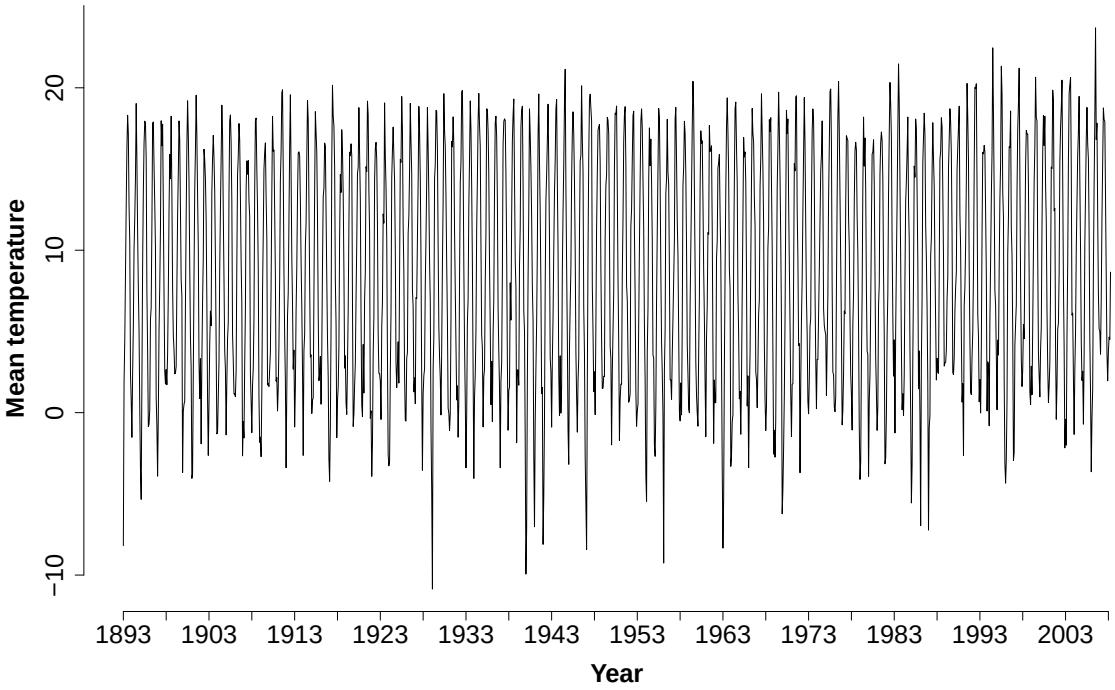

Figure 4.15: Original mean temperature time series.

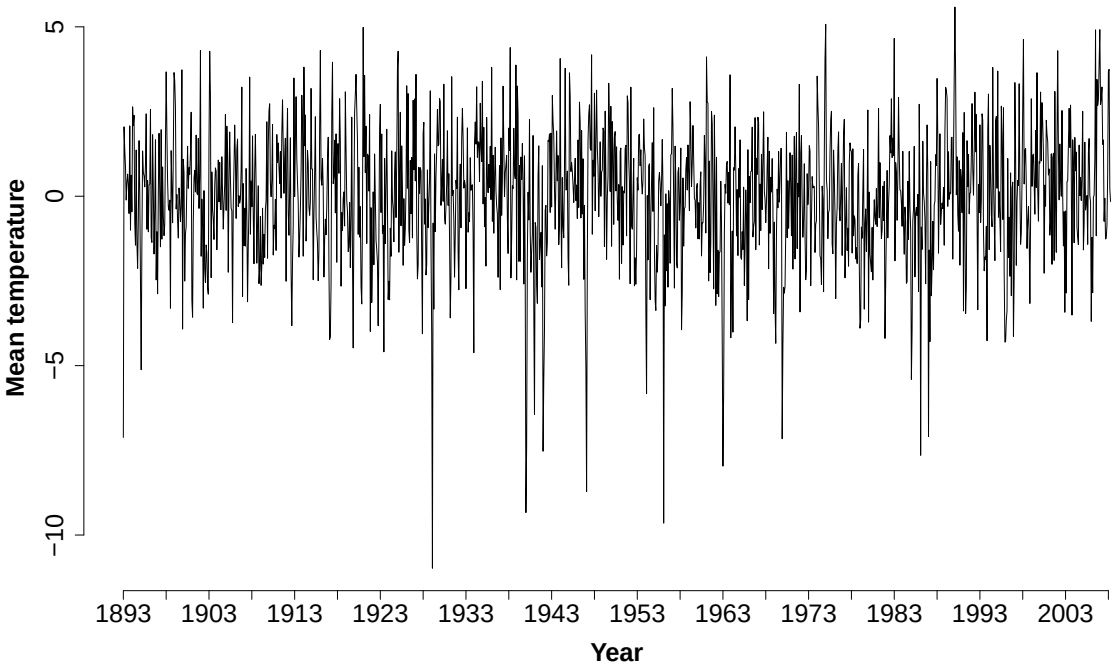

Figure 4.16: Detrended and deseasonalized mean temperature time series.

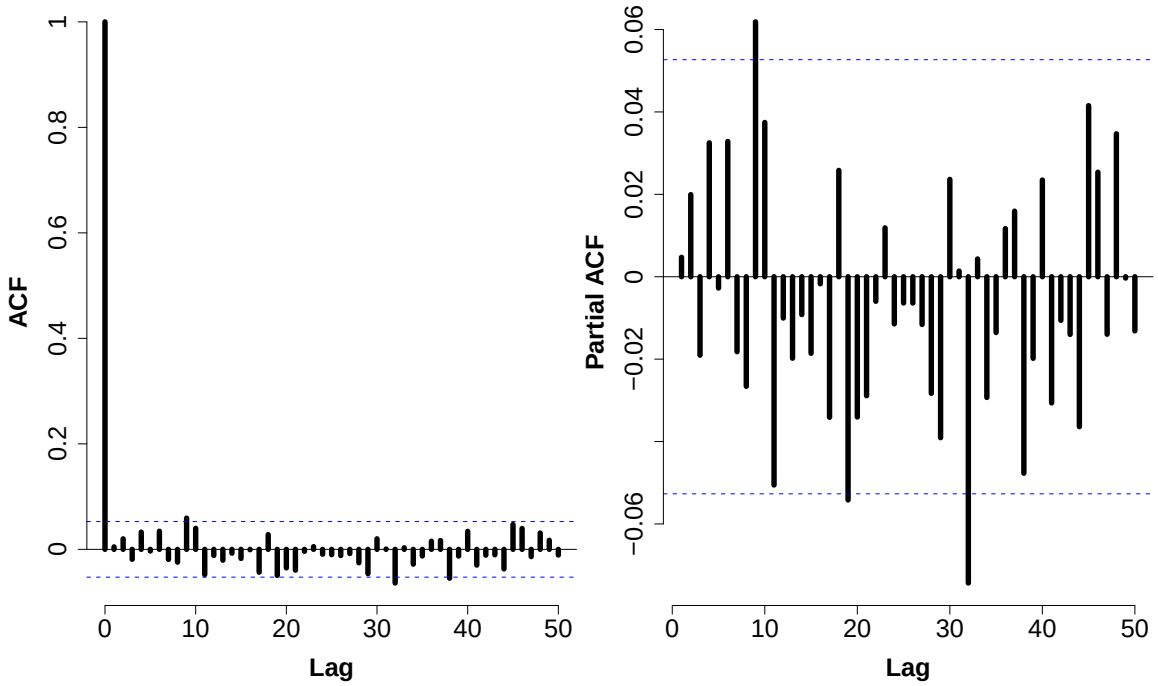

Figure 4.17: Autocorrelation (left) and partial autocorrelation function (right) of the detrended and deseasonalized total rainfall time series.

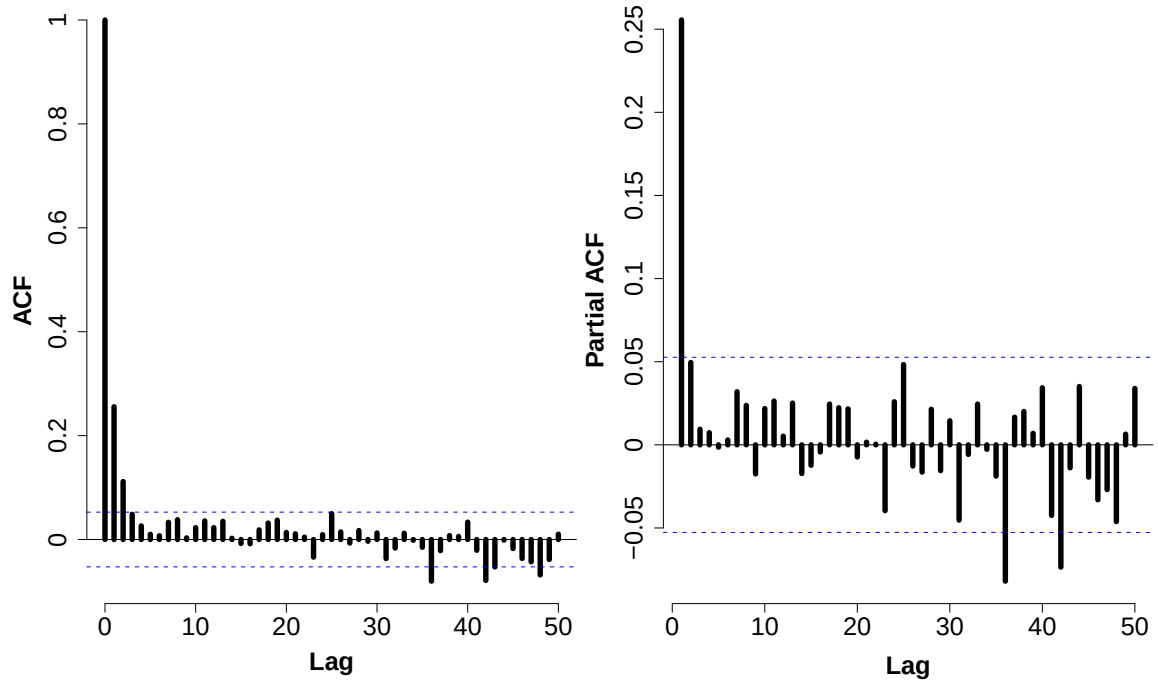

Figure 4.18: Autocorrelation (left) and partial autocorrelation function (right) of the detrended and deseasonalized mean temperature time series.

Table 4.1: p–values for the total rainfall time series (in percent)

$\widetilde{k}$	12	24	60	120	240	360
U^o	6.4	40.9	11.7	18.3	11.1	6.1
L^r	9.3	21.3	32.4	26.8	38.7	7.9
T_1	4.2	34.9	14.2	7.9	14.8	9.8
T_2	4.3	31.8	3.3	11.9	12.8	7.4
T_3	2.3	23.7	12.9	15.7	17.8	4.6
T_4	1.9	22.5	6.0	7.8	17.6	7.5
S	17.2	12.8	28.1	25.6	24.1	9.1
R_1	19.4	15.7	33.2	39.2	37.5	13.0
R_2	26.7	19.2	36.3	42.2	33.1	26.5
R_3	44.0	38.6	57.0	58.9	45.5	11.1
R_4	48.7	44.8	63.4	61.8	41.2	20.5
N_1	8.2	35.6	32.4	18.6	5.1	5.8
N_2	4.6	5.1	58.4	61.7	49.1	20.0
N_3	61.1	61.1	46.1	46.1	46.1	14.6

For the total rainfall time series the record tests T_1, T_2, T_3 and T_4 with $\widetilde{k} = 12$ detect a monotone trend at a significance level of $\alpha = 0.05$. From the rank tests only N_2 finds a monotone trend at this α. Using a larger splitting factor we only find a monotone trend with T_2 for $\widetilde{k} = 60$. Of course we need to keep in mind that we perform multiple testing and thus expect about four significant test statistics among the more than 80 tests performed here even if there is no trend at all.

Table 4.2: p–values for the mean temperature time series (in percent)

$\widetilde{k}$	12	24	60	120	240	360
U^o	0.00	0.00	0.00	0.00	0.00	0.00
L^r	0.00	0.03	0.01	0.00	0.00	0.00
T_1	0.00	0.00	0.00	0.00	0.00	0.00
T_2	0.00	0.00	0.00	0.00	0.00	0.00
T_3	0.00	0.00	0.00	0.00	0.00	0.00
T_4	0.00	0.00	0.00	0.00	0.00	0.00
S	0.00	0.00	0.00	0.00	0.00	0.00
R_1	0.00	0.00	0.00	0.00	0.00	0.00
R_2	0.00	0.00	0.00	0.00	0.00	0.00
R_3	0.00	0.00	0.00	0.00	0.00	0.00
R_4	0.00	0.00	0.00	0.00	0.00	0.00
N_1	97.42	13.40	5.04	21.07	0.05	0.06
N_2	0.00	0.00	0.00	0.00	0.00	0.00
N_3	0.00	0.00	0.00	0.00	0.00	0.00

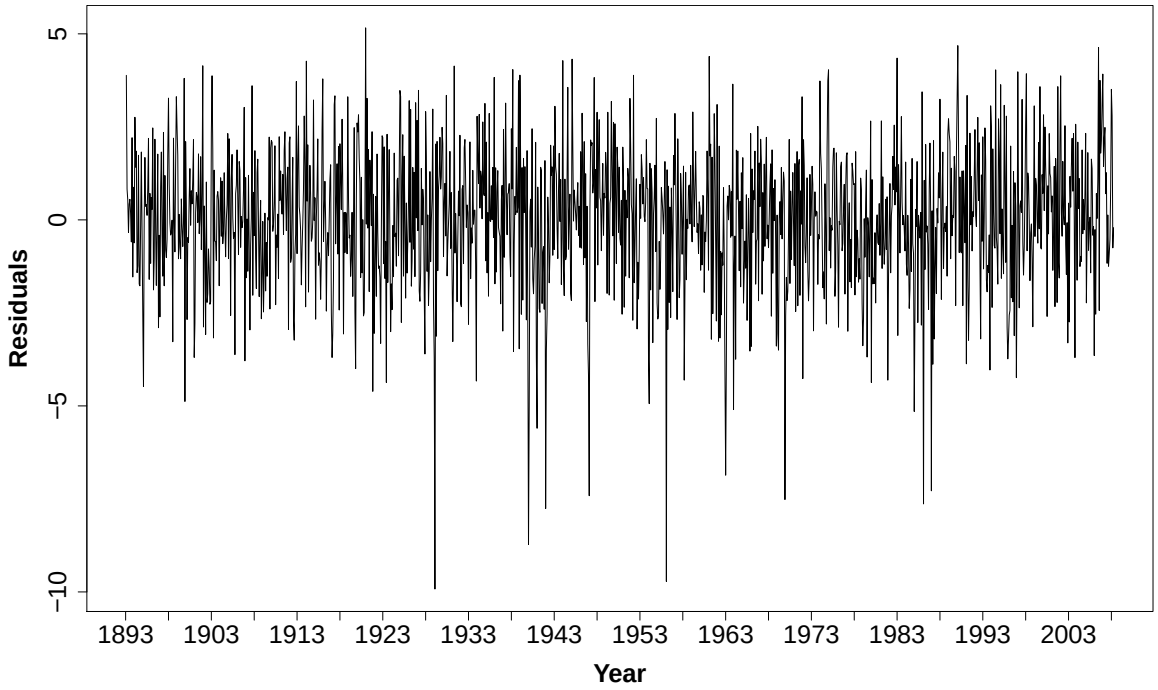

Figure 4.19: Residuals of the temperature time series obtained from fitting an AR(1) model to the deseasonalized temperature time series.

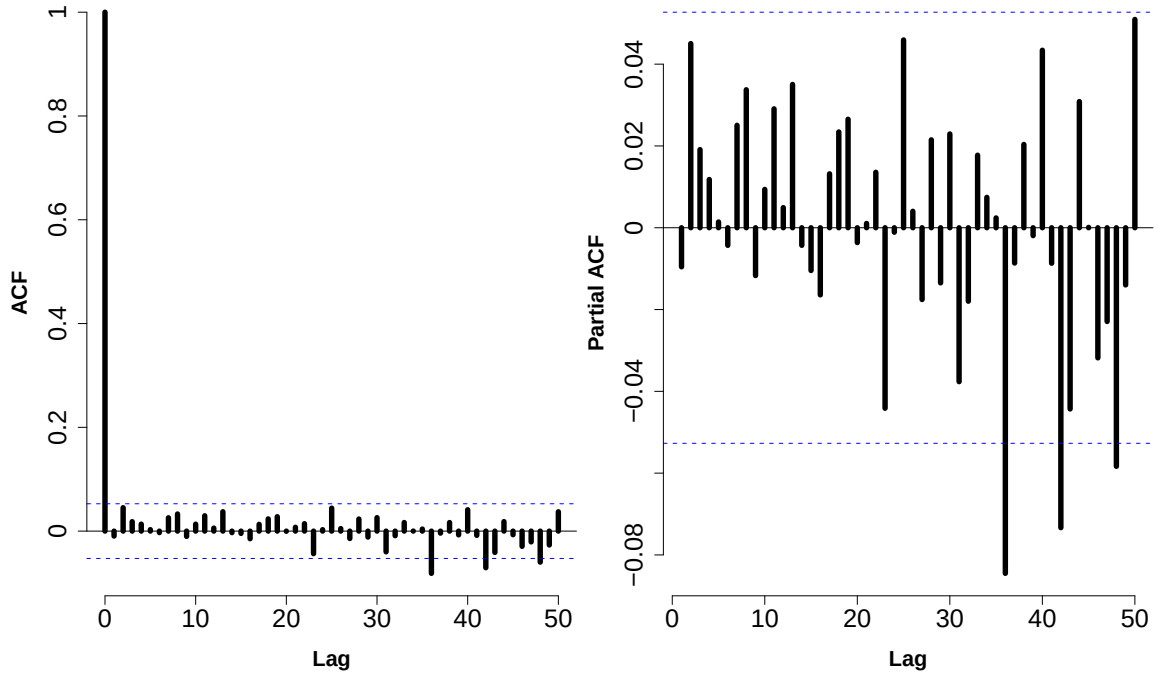

Figure 4.20: ACF (left) and PACF (right) of the AR(1) residuals of the deseasonalized temperature series.

Table 4.3: p–values for the residual temperature time series (in percent)

$\widetilde{k}$	12	24	60	120	240	360
U^o	0.30	0.19	0.07	0.07	0.00	0.15
L^r	2.77	0.41	0.13	0.93	0.24	0.09
T_1	0.01	0.01	0.00	0.00	0.00	0.01
T_2	0.44	0.07	0.05	0.08	0.00	0.12
T_3	0.05	0.01	0.00	0.01	0.00	0.03
T_4	0.02	0.00	0.00	0.01	0.00	0.01
S	0.00	0.00	0.00	0.00	0.00	0.01
R_1	0.00	0.00	0.00	0.00	0.00	0.00
R_2	0.00	0.00	0.00	0.00	0.00	0.02
R_3	0.00	0.00	0.00	0.00	0.00	0.00
R_4	0.00	0.00	0.00	0.00	0.00	0.01
N_1	93.10	23.01	11.80	53.56	0.10	1.91
N_2	0.00	0.00	0.00	0.00	0.01	0.01
N_3	0.01	0.03	0.00	0.00	0.00	0.00

All tests except N_1 detect a monotone trend in the temperature time series for all splittings $\widetilde{k}$. The statistic N_1 only detects a monotone trend, if $\widetilde{k}$ is large. But as all tests need the assumption of independence, the results of Table 4.2 can not be interpreted as p–values of unbiased tests. This is why we deseasonalize the temperature time series and fit an AR(1)–Model to the deseasonalized series by maximum likelihood. If the data generating mechanism is an AR(1) process with uncorrelated innovations, then the residuals of the fitted AR(1) model are asymptotically uncorrelated. The residuals are even asymptotically independent, if the innovations are i.i.d. The residuals are asymptotically normal, if the innovations are normally distributed (see Section 5.3 of Brockwell 2002). Looking at the plot of the scaled residual time series in Fig. 4.19 and its ACF in Fig. 4.20, we do not find significant autocorrelations between the residuals. However, the residuals do not seem to be identically normally distributed, as we can find some outliers in the residual plot. Table 4.3 shows the p–values of the record and rank tests for the residuals. We find mostly larger p–values than in Table 4.2, but again all tests except N_1 detect a positive monotone trend at $\alpha = 0.05$, what confirms the previous findings.

4.5 Conclusions

We have considered nonparametric tests for trend detection in time series. We have not found large differences between the power of the different tests. All tests based on records or ranks react sensitive to autocorrelations. Our results confirm findings by Diersen and Trenkler that T_3 can be recommended among the record tests because of its good power and its simplicity. Robustness of T_3 against autocorrelation can be achieved for the price of a somewhat reduced power by choosing a large splitting

factor $\widetilde{k}$. However, even higher power can be achieved by applying a nonparametric rank test like the seasonal Kendall–Test S or the Spearman–Test R_1 with a small $\widetilde{k}$, even though for the price of a higher sensitivity against positive autocorrelation. The power of all rank tests except N_1 gets smaller, if a larger splitting factor is used. For N_1 a larger splitting factor enlarges the power, but N_1 is not recommended for use, as even with a large splitting factor it is less powerful than the other tests. From the rank tests the test N_2 is more robust against autocorrelations below 0.6 than the other tests, if only three observations are taken in each block. Another possibility to reduce the sensitivity to autocorrelation is to fit a low order AR model and consider the AR residuals. We have found a significant trend in the time series of the monthly mean temperature in Potsdam both when using the original data and the AR(1) residuals. Since in the plot of the scaled residuals for this series we find some outliers, another interesting question for further research is the robustness of the several tests against atypical observations.

List of Tables

List of Figures

Chapter 2

Chapter 3

Chapter 4

References

Aiyar, R.J., Guillier, C.L., Albers, W. (1979): Asymptotic relative efficiencies of rank tests for trend alternatives, *Journal of the American Statistical Association* **74**, 227–231.

Benhenni, K., Ferraty, F., Rachdi, M., Vieu, P. (2007): Local smoothing regression functional data, *Computational Statistics* **22**, 353–369.

Bernholt, T., Fried, R. (2003): Computing the update of the repeated median regression line in linear time, *Information Processing Letters* **88**, 111–117.

Bernholt, T., Fried, R., Gather, U., Wegener, I. (2006): Modified repeated median filters, *Statistics and Computing* **16**, 177–192.

Bianco, A., Boente, G. (2006): Robust estimators under semi-parametric partly linear autoregression: asymptotic behaviour and bandwidth selection, *Journal of Time Series Analysis* **28**, 274–306.

Boente, G., Rodriguez, D. (2008): Robust bandwidth selection in semi-parametric partly linear regression models: Monte Carlo study and influential analysis, *Computational Statistics and Data Analysis* **52**, 2808–2828.

Brockwell, P.J., Davis, R.A. (2002): *Introduction to time series and forecasting*, Second Edition, Springer, New York.

Cantoni, E., Ronchetti, E. (2001): Resistant selection of the smoothing parameter for smoothing splines, *Statistics and Computing* **11**, 141–146.

Cleveland, W.S. (1979): Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association* **74**, 829–836.

Cobb, G. (1978): The problem of the Nile: conditional solution to a change-point problem, *Biometrika* **65**, 243–251.

Cox, D.R. Stuart, A. (1955): Some quick sign tests for trend in location and dispersion, *Biometrika* **42**, 80–95.

Croux, C., Rousseeuw, P.J., Hossjer, O. (1994): Generalized S-estimators, *Journal of the American Statistical Association* **89**, 1271–1281.

Daniels, H.E. (1950): Rank correlation and population models, *Journal of the Royal Statistical Society, Series B* **12**, 171–181.

Davies, P.L., Fried, R., Gather, U. (2004): Robust signal extraction for on-line monitoring data. *Journal of Statistical Planning and Inference* **122**, 65–78.

Diersen, J., Trenkler, G. (1996): Records tests for trend in location, *Statistics* **28**, 1–12.

Diersen, J., Trenkler, G. (2001): Weighted record tests for splitted series of observations, In: Kunert, J., Trenkler, G. (eds.): *Mathematical statistics with applications in biometry, Festschrift in Honour of Prof. Dr. Siegfried Schach*, Josef Eul, Lohmar, DE, 163–178.

Donoho, D.L., Huber, P.J. (1983): The notion of breakdown point. In: Bickel, P.J., Doksum K., Hodges J.L. (eds): *A Festschrift for Erich Lehmann*, Wadsworth, Belmont, CA, 157–184.

Donoho, D.L., Johnstone, I.M. (1994): Ideal spatial adaptation by wavelet shrinkage, *Biometrika* **81**, 425–455.

Fan, J., Hall, P., Martin, M.A., Patil, P. (1996): On local smoothing of nonparametric curve estimators, *Journal of the American Statistical Association* **91**, 258–266.

Fearnhead, P., Clifford, P. (2003): On-line inference for hidden markov models via particle filters, *Journal of the Royal Statistical Society, Series B* **65**, 887–889.

Foster, F.G., Stuart, A. (1954): Distribution-free tests in time-series based on the breaking of records, *Journal of the Royal Statistical Society, Series B* **16**, 1–22.

Francisco-Fernandez, M., Vilar-Fernandez, J.M. (2005): Bandwidth selection for the local polynomial estimator under dependence: a simulation study, *Computational Statistics* **20**, 539–558.

Fried, R. (2007): On the robust detection of edges in time series filtering, *Computational Statistics and Data Analysis* **52**, 1063–1074.

Fried, R., Bernholt, T., Gather, U. (2006): Repeated median and hybrid filters, *Computational Statistics and Data Analysis* **50**, 2313–2338.

Fried, R., Dehling, H. (2011): Robust nonparametric tests for the two sample location problem, *Statistical Methods and Applications* **20**, 409–422.

Gather, U., Schettlinger, K., Fried, R. (2006): Online signal extraction by robust linear regression, *Computational Statistics* **21**, 33–51.

Gijbels, I., Goderniaux, A.C. (2004): Bandwidth selection for changepoint estimation, *Technometrics* **46**, 76–86.

Gijbels, I., Goderniaux, A.C. (2004): Bootstrap test for change-points in nonparametric regression, *Journal of Nonparametric Statistics* **16**, 591–611.

Gijbels, I., Lambert, A., Qiu, P. (2007): Jump-preserving regression and smoothing using local linear fitting: a compromise, *The Annals of the Institute of Statistical Mathematics* **59**, 235–272.

Haerdle, W. (1984): How to determine the bandwidth of nonlinear smoothers in practice?, In: Franke J., Haerdle W., Martin, D. (eds): *Lecture Notes in Statistics 26*, Springer, Heidelberg, DE, 163–184.

Haerdle, W. (2002): *Applied nonparametric regression*, Cambridge University Press, Edinburgh.

Haerdle, W., Hall, P., Marron, J.S. (1988): How far are automatically chosen regression smoothing parameters from their optimum?, *Journal of the American Statistical Association* **83**, 86–95.

Haerdle, W., Marron, J.S. (1985): Optimal bandwidth selection in nonparametric regression function estimation, *The Annals of Statistics* **13**, 1465–1481.

Heinonen, P., Neuvo, Y. (1987): FIR-median hybrid filters, *IEEE Transactions of Acoustics, Speech and Signal Processing* **35**, 832–838.

Hillebrand, M., Mueller, C.H. (2006): On consistency of redescending M-kernel smoothers, *Metrika* **63**, 71–90.

Hirsch, R.M., Slack, J.R. Smith, R.A. (1982): Techniques of trend analysis for monthly water quality data, *Water Resources Research* **18**, 107–121.

Hirsch, R.M., Slack J.R. (1984): A nonparametric trend test for seasonal data with serial dependence, *Water Resources Research* **20**, 727–732.

Hodges, J.L., Lehmann, E.L. (1963): Estimates of location based on rank tests, *The Annals of Mathematical Statistics* **34**, 598–611.

Holm, S. (1979): A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics* **6**, 65–70.

Kendall, M.G. (1938): A new measure of rank correlation, *Biometrika* **30**, 81–93.

Kendall, M.G., Gibbons, J.D. (1990): *Rank correlation methods*, Arnold, London.

Kerkyacharian, K., Lepski, O., Picard, D. (2001): Nonlinear estimation in anisotropic multi-index denoising, *Probability Theory and Related Fields* **121**, 137–170.

Lafferty, J., Wasserman, L. (2008): Rodeo: Sparse, greedy nonparametric regression, *The Annals of Statistics* **36**, 28–63.

Lee, J.S., Cox, D.D. (2010): Robust smoothing: Smoothing parameter selection and applications to fluorescence spectroscopy, *Computational Statistics and Data Analysis* **54**, 3131–3143.

Lee, Y.H., Kassam, S.A. (1985): Generalized median filters and related nonlinear filtering techniques, *IEEE Transactions of Acoustics, Speech and Signal Processing* **33**, 672–683.

Leung, D.H.Y. (2005): Cross-validation in nonparametric smoothing with outliers, *The Annals of Statistics* **33**, 2291–2310.

Leung, D.H.Y., Marriott, F.H.C., Wu, E.K.H. (1993): Bandwidth selection in robust smoothing, *Journal of Nonparametric Statistics* **2**, 333–339.

Maechler, M. (1989): Parametric smoothing quality in nonparametric regression: Shape control by penalizing inflection points, Ph.d. thesis, no 8920, ETH Zuerich, Statistik, ETH-Zentrum, CH-8092 Zurich, Switzerland.

Mann, H.B. (1945): Non-parametric tests against trend, *Econometrica* **13**, 245–259.

Maronna, R.A., Martin, R.D., Yohai, V.J. (2006): *Robust statistics*, John Wiley & Sons Ltd., Chichester.

Maronna, R.A., Zamar, R.H. (2002): Robust estimates of location and dispersion of high-dimensional datasets, *Technometrics* **44**, 307–317.

Moore, G.H., Wallis, W.A. (1943): Time series significance tests based on signs of differences, *Journal of the American Statistical Association* **38**, 153–164.

Morell, O., Bernholt, T., Fried, R., Kunert, J., Nunkesser, R. (2008): An evolutionary algorithm for LTS-regression: a comparative study, In: Brito, P. (Eds.): *COMPSTAT 2008: Proceedings in Computational Statistics, Vol. II*, Physica, Heidelberg, DE, 585–593.

Morell, O., Fried, R. (2009): On nonparametric tests for trend detection in seasonal time series, In: Schipp, B., Krämer, W. (Eds): *Statistical Inference, Econometric Analysis and Matrix Algebra, Festschrift in Honour of Götz Trenkler*, Physica, Heidelberg, DE, 19–40.

Morell, O., Otto, D., Fried, R. (2012): On robust cross-validation for nonparametric smoothing, to appear in *Computational Statistics*.

Mueller, C.H. (2002): Robust estimators for estimating discontinuous functions, *Metrika* **55**, 99–109.

Nadaraya, E.A. (1964): On estimating regression, *Theory of Probability and Applications* **9**, 141–142.

Noether, G.E. (1955): On a theorem of Pitman, *The Annals of Mathematical Statistics* **26**, 64–68.

Nunkesser, R., Morell, O. (2010): An evolutionary algorithm for robust regression, *Computational Statistics and Data Analysis* **54**, 3242–3248.

O Ruanaidh, J.J.K., Fitzgerald, W.J. (1996): *Numerical bayesian methods applied to signal processing*, Springer, New York.

Qiu, P. (1994): Estimation of the number of jumps of the jump regression function, *Communications in Statistics - Theory and Methods* **23**, 2141–2155.

Qiu, P., Yandell, B. (1998): A local polynomial jump-detection algorithm in nonparametric regression, *Technometrics* **40**, 141–152.

R Development Core Team (2011): R: A language and environment for statistical computing, *R Foundation for Statistical Computing*, Vienna, Austria, ISBN 3-900051-07-0, URL http://www.R-project.org.

Rousseeuw, P.J. (1984): Least median of squares regression, *Journal of the American Statistical Association* **79**, 871–880.

Rousseeuw, P.J., Croux, C. (1993): Alternatives to the median absolute deviation, *Journal of the American Statistical Association* **88**, 1273–1283.

Schmidt, G., Mattern, R., Schueler, F. (1981): Biomechanical investigation to determine physical and traumatological differentiation criteria for the maximum load capacity of head and vertebral column with and without protective helmet under effects of impact, *EEC Research Program on Biomechanics of Impacts*, Final Report Phase III, Project 65, Institut fuer Rechtsmedizin, Universitaet Heidelberg, West Germany.

Serneels, S., Filzmoser, P., Croux, C., Van Espen, P.J. (2005): Robust continuum regression, *Chemometrics and Intelligent Laboratory Systems* **76**, 197–204.

Shibata, R. (1981): An optimal selection of regression variables, *Biometrika* **68**, 45–54.

Siegel, A.F. (1982): Robust regression using repeated medians, *Biometrika* **69**, 242–244.

Silverman, B.W. (2005): *EbayesThresh: Empirical Bayes thresholding and related methods*, R package version 1.3.0., URL: http://www.bernardsilverman.com.

Spearman, C.E. (1904): The proof and measurement of association between two things, *American Journal of Psychology* **15**, 72–101.

Tukey, J.W., McLaughin, D.H. (1963): Less vulnerable confidence and significance procedures for location based on a single sample: trimming / winsorization 1, *Sankhya, Series A* **25**, 331–352.

Wang, F.T., Scott, D.W. (1994): The L_1 method for robust nonparametric regression, *Journal of the American Statistical Association* **89**, 65–76.

Watson, G.S. (1964): Smooth regression analysis, *Sankhya, Series A* **26**, 359–372.

Wu, J.S., Chu, C.K. (1993): Kernel-type estimators of jump points and values of a regression function, *The Annals of Statistics* **21**, 1545–1566.

Yang, Y., Zheng, Z. (1992): Asymptotic properties for cross-validated nearest neighbor median estimators in nonparametric regression: the L_1-view, In: Jiang, Z., Yan, S., Cheng, P., Wu, R. (eds): *Probability and Statistics*, World Scientific, SG, 242–257.

Ylvisaker, D. (1977): Test resistance, *Journal of the American Statistical Association* **72**, 551–556.

Yuen, K.K., Dixon, W.J. (1973): The approximate behaviour and performance of the two sample trimmed t, *Biometrika* **60**, 369–374.

Zhang, X., Brooks, R.D., King, M.L. (2009): A Bayesian approach to bandwidth selection for multivariate kernel regression with an application to state-price density estimation, *Journal of Econometrics* **153**, 21–32.

Zheng, Z., Yang, Y. (1998): Cross-validation and median criterion, *Statistica Sinica* **8**, 907–921.

EIDESSTATTLICHE ERKLÄRUNG

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

Dortmund, den 13.08.2012

Oliver Morell