On the Robust Detection of Edges in Time Series Filtering

Roland Fried

Department of Statistics, University of Dortmund, 44221 Dortmund, Germany

Abstract: Abrupt shifts in the level of a time series represent important information and should be preserved in statistical signal extraction. We investigate rules for detecting level shifts that are resistant to outliers and which work with only a short time delay. The properties of robustified versions of the t-test for two independent samples and its non-parametric alternatives are elaborated under different types of noise. Trimmed t-tests, median comparisons, robustified rank and ANOVA tests based on robust scale estimators are compared.

Keywords: Time series filtering, Jumps, Outliers, Test resistance

1 Introduction

An important task in statistical signal extraction is the detection of abrupt shifts (also called step changes, edges or jumps). This task is complicated by the presence of outliers, since these can easily be confused with shifts, particularly in an online analysis, when only a short time delay is permitted. An abundance of rules for level-shift detection has been suggested in the literature, but many rules are unable to distinguish between outliers and shifts. This distinction is the goal here.

We use a components model for the observed time series (y_t):

$$y_t = \mu_t + u_t + v_t,\ t \in \mathbb{Z}. \tag{1}$$

The level μ_t of the time series (the underlying signal) is assumed to vary smoothly over time with a few abrupt shifts. The noise is assumed to consist of an ordinary random disturbance u_t, which is distributed symmetrically with a zero mean and a variance σ_t^2 (which may be time-varying), together with an intermittent outlier component v_t, which is of an impulsive (spiky) nature. The spiky noise is zero most of the time, but, occasionally, it takes large absolute values.

We investigate rules for shift detection which are straightforward when using a moving-window approach for signal extraction. Moving averages approximate μ_t efficiently in case where u_t is a Gaussian noise, but they are sensitive to impulsive noise and they blur level shifts. Standard median filters (also called running medians) perform better in these respects (Tukey, 1977, Nieminem, Neuvo and Mitra, 1989). They approximate the signal μ_t in the centre of a time window $(y_{t-k}, \ldots, y_{t+k})$ of width $n = 2k+1$ by the median of the observations,

$$StM(y_t) \;=\; \tilde{\mu}_t = med(y_{t-k}, \ldots, y_{t+k}),\; t \in \mathbb{Z}.$$

As a compromise between the mean and the median, we can calculate an α-trimmed mean, which is the mean of the remaining observations after deleting the largest and the smallest $\lfloor \alpha n \rfloor$ values in the window, where $\lfloor \alpha n \rfloor$ denotes the integer part of αn. In general, every location estimator is a candidate for an approximation of the level at the centre of the window. Rules for level-shift detection arise from the many filtering procedures based on moving windows.

In this paper, we compare new and existing rules for automatic level-shift detection, which are based on moving-window operations and which do not need a global parametric model of the data. Robustified exponentially-weighted moving average (EWMA, Cypra, 1992) or CUSUM charts (Zeileis, 2005) are not considered here, since they are more difficult to handle than moving-window techniques if we require the procedure to resist a predetermined number of outliers (Imhoff et al., 2002). Also, they react to other changes such as drifts, whereas we aim at a procedure that indicates only level shifts.

The paper is organised as follows. Section 2 presents rules for robust shift detection in time series. Section 3 compares the rules analytically and via simulations. Section 4 presents an application, after which we summarise the results.

2 Shift detection

We assume that an ideal shift of height $\delta \in \mathbb{R}$ occurs between time points t and $t+1$, and that it may be accompanied by a simultaneous change of the variance of the ordinary noise component u_t:

$$
\begin{aligned}
y_{t+j} &= \begin{cases} \mu + u_{t+j} + v_{t+j}, & j = \ldots, -1, 0, \\ \mu + \delta + u_{t+j} + v_{t+j}, & j = 1, 2, \ldots, \end{cases} \qquad (2) \\
u_{t+j} &\sim \begin{cases} N(0, \sigma^2), & j = \ldots, -1, 0, \\ N(0, \kappa\sigma^2), & j = 1, 2, \ldots \ . \end{cases}
\end{aligned}
$$

To detect a positive (negative) shift immediately after time t, we test $H_0 : \delta = 0$ versus $H_1^+ : \delta > 0$ ($H_1^- : \delta < 0$). The variance of the noise may be unaffected, with $\kappa = 1$, or it may change, with $\kappa > 0$ taking some arbitrary positive value. In what follows, some detection schemes are presented for testing H_0 versus H_1^+ and / or H_1^-. These schemes assume that the noise constitutes a serially independent sequence, but the simulations reported below show that moderate autocorrelations do not have large effects.

2.1 Gradient detection schemes

Gradient schemes for detecting whether a level shift has occurred immediately after time t compare two level estimates $\hat{y}_{t-}$ and $\hat{y}_{t+}$ calculated from windows $(y_{t-h+1}, \ldots, y_t)$ and $(y_{t+1}, \ldots, y_{t+k})$ of widths h and k, which may differ. In general, a shift is detected if

$$
\frac{|\hat{y}_{t+} - \hat{y}_{t-}|}{\hat{\tau}_t} > q \ ,
$$

where $\hat{\tau}_t$ is a standardisation specified below and q is an appropriate threshold.

The ordinary two-sample t-test is obtained by setting $\hat{y}_{t+}$ and $\hat{y}_{t-}$ equal to the averages $\overline{y}_{t+}$ and $\overline{y}_{t-}$ of the data in the two windows (Stein and Fowlow, 1985) and by assuming that the variance σ^2_{t+j} is constant within $\{t-h+1, \ldots, t+k\}$. Thus

$$
\begin{aligned}
T &= \frac{\overline{y}_{t+} - \overline{y}_{t-}}{\hat{s}_t \sqrt{1/h + 1/k}} \ , \qquad (3) \\
\hat{s}_t^2 &= \frac{1}{n-2} \left[\sum_{i=0}^{h-1} (y_{t-i} - \overline{y}_{t-})^2 + \sum_{j=1}^{k} (y_{t+j} - \overline{y}_{t+})^2 \right] \ ,
\end{aligned}
$$

where $n = h + k$ is the total number of observations in the two windows. If the noise is assumed to be Gaussian, then the threshold value q is a quantile of the t-distribution with $n - 2$ degrees of freedom.

As alternatives to the ordinary means, we can use the α-trimmed means $\hat{y}_{t+} = \overline{y}_{t+}^{(\alpha)}$ and $\hat{y}_{t-} = \overline{y}_{t-}^{(\alpha)}$. This enhances the robustness at the cost of a small loss of efficiency in the case of Gaussian noise (Yuen, 1974, Bovik and Munson, 1986, Hou and Koh, 2003, Fried, 2007). A trimmed t-test can be constructed by standardising $|\overline{y}_{t+}^{(\alpha)} - \overline{y}_{t-}^{(\alpha)}|$ using the α-winsorised variance of the residuals $y_{t-h+1} - \overline{y}_{t-}^{(\alpha)}, \ldots, y_t - \overline{y}_{t-}^{(\alpha)}, y_{t+1} - \overline{y}_{t+}^{(\alpha)}, \ldots, y_{t+k} - \overline{y}_{t+}^{(\alpha)}$. This is the empirical variance of a modified set of residuals obtained by replacing the $\lfloor \alpha n \rfloor$ largest residuals by the next largest residual and by replacing the $\lfloor \alpha n \rfloor$ smallest residuals by the next smallest one. A winsorised variance is a robust estimator in the presence of outliers.

The median, which is the 50%-trimmed mean, has been suggested for edge detection in images with heavy-tailed noise (Bovik and Munson, 1986, Hwang and Haddad, 1994). In image analysis, the noise variance is often regarded as globally constant, in which case very good estimates of it are available. However, we are concerned with time series and we wish to make minimal assumptions. If the noise distribution possesses a density f with a zero median, then the median of k independent observations is approximated, with increasing accuracy as $k \to \infty$, by a normal distribution with variance $1/(4kf^2(0))$ (e.g. Stigler, 1973). Assuming that the noise is Gaussian with variances of σ^2 and $\kappa\sigma^2$ in the left and the right-hand windows, respectively, then the difference of the corresponding medians $\tilde{y}_{t-}$ and $\tilde{y}_{t+}$ has a zero mean and an asymptotic variance of $0.5\pi(\sigma^2/h + \kappa\sigma^2/k)$ under the null hypothesis of no shift. The following test statistics are asymptotically standard normal under the null:

$$\frac{\tilde{y}_{t+} - \tilde{y}_{t-}}{\sqrt{0.5\pi\hat{\sigma}_t^2(1/h + 1/k))}}\ , \tag{4}$$

where we assume that $\kappa = 1$, i.e. identical variances in both windows, and

$$\frac{\tilde{y}_{t+} - \tilde{y}_{t-}}{\sqrt{0.5\pi(\hat{\sigma}_{t-}^2/h + \hat{\sigma}_{t+}^2/k)}}\ , \tag{5}$$

where we assume that $\kappa \neq 1$ and where $\hat{\sigma}_t^2$, $\hat{\sigma}_{t-}^2$ and $\hat{\sigma}_{t+}^2$ are consistent robust variance estimators obtained from the whole window, and from the left and the right windows, respectively.

The α-winsorised variance can be applied for standardising α-trimmed means only if α is substantially less than 50%, but it is not appropriate for the median.

Instead, we can use a highly robust scale estimator such as the median absolute deviation about the median (MAD). Assuming a constant noise variance σ^2, we can combine the differences obtained in both windows to form

$$\hat{\sigma}_t^{(M)} = c_n^{(M)} med(|y_{t-h+1} - \tilde{y}_{t-}|, \ldots, |y_t - \tilde{y}_{t-}|, |y_{t+1} - \tilde{y}_{t+}|, \ldots, |y_{t+k} - \tilde{y}_{t+}|). \quad (6)$$

Here, $c_n^{(M)}$ is a finite-sample correction factor, which becomes 1.4826 for very large n. Otherwise, if the noise variances differ, we can use two MADs calculated from the left and the right windows,

$$\begin{aligned} \hat{\sigma}_{t-}^{(M)} &= c_h^{(M)} med(|y_{t-h+1} - \tilde{y}_{t-}|, \ldots, |y_t - \tilde{y}_{t-}|) , \\ \hat{\sigma}_{t+}^{(M)} &= c_k^{(M)} med(|y_{t+1} - \tilde{y}_{t+}|, \ldots, |y_{t+k} - \tilde{y}_{t+}|) . \end{aligned} \quad (7)$$

Some alternatives to the MAD have been introduced. Scale estimators measuring the variability via the differences between the observations do not need an estimator of central location. This can be an advantage, since location estimators become biased in the vicinity of a level shift. The estimators described in the remainder of this section possess an asymptotic explosion breakdown point of 50% like the MAD, meaning that the increase of the estimate caused by replacing less than half of the data in a given sample by arbitrary values is always bounded.

The LSH (length of the shortest half) estimator of variability (Grübel, 1988, Rousseeuw and Leroy, 1988) is represented, in our notation, by

$$\hat{\sigma}_t^{(L)} = c_n^{(L)} min(z_{(n)} - z_{(n-m)}, \ldots, z_{(m+1)} - z_{(1)}) . \quad (8)$$

Here, $m = \lfloor (n+1)/2 \rfloor$, and the $z_{(i)}$ are the differences $y_{t-h+1} - \tilde{y}_{t-}, \ldots, y_t - \tilde{y}_{t-}, y_{t+1} - \tilde{y}_{t+}, \ldots, y_{t+k} - \tilde{y}_{t+}$ ordered according to their size. Again, $c_n^{(L)}$ is a correction factor, designed to achieve unbiasedness in a Gaussian sample of size n.

The Sn statistic of Rousseeuw and Croux (1993) is another means of estimating the variability, which is represented, in our notation, by

$$\hat{\sigma}_t^{(S)} = c_n^{(S)} med_i med_{j \neq i} |z_i - z_j| . \quad (9)$$

Finally, we can use the Qn statistic, also suggested by Rousseeuw and Croux (1993). On the assumption of a constant variance, it is calculated from the full

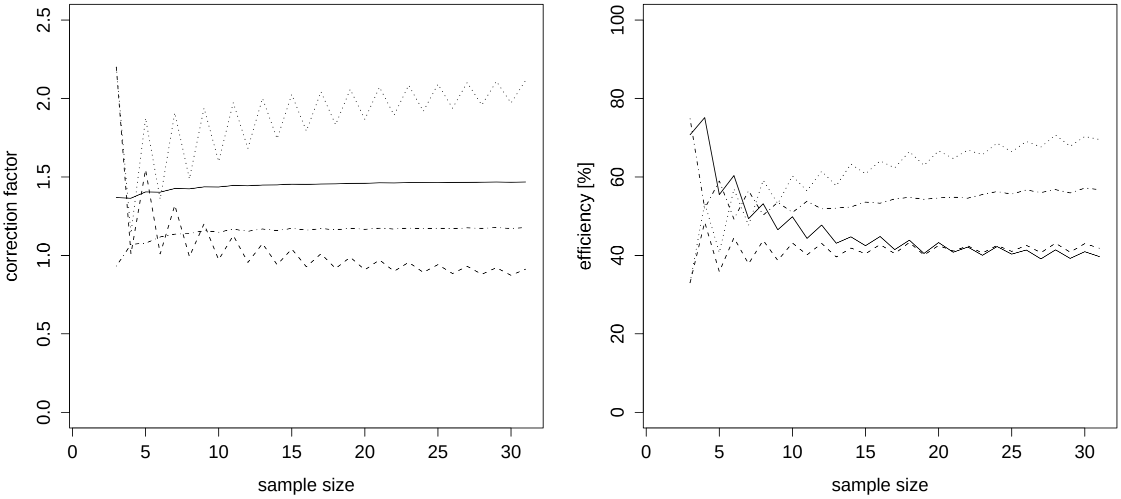

Figure 1: Finite-sample correction factors (left) and efficiencies (right) of the MAD (solid line), LSH (dashed), Sn (dash-dot) and Qn (dotted) for different sample sizes.

window via the formula

$$\hat{\sigma}_t^{(Q)} = c_n^{(Q)}(|z_i - z_j|, 1 \leq i < j \leq n)_{\left(\binom{\lfloor n/2 \rfloor + 1}{2}\right)} , \tag{10}$$

using the $\binom{\lfloor n/2 \rfloor + 1}{2}$-th smallest of the $\binom{n}{2}$ absolute pairwise differences $|z_i - z_j|$, i.e. approximately the first quartile of all pairwise differences. Of course, we can also estimate the variability in the two windows separately. For more information on these estimators see Gather and Fried (2003).

Fig. 1 shows the finite-sample correction factors for the different scale estimators as a function of the underlying sample size as well as the finite-sample efficiencies relatively to the empirical standard deviation as measured by the percentage of the mean square error (MSE) under Gaussian noise.

2.2 Tests based on estimates of the local variability

Comparison of the levels of two time windows can be treated within the framework of analysis of variance (ANOVA). The ANOVA F-test compares the variability between the groups to that within the groups. In the case of two groups, it is just the square of the ordinary two-sample t-test presented in Section 2.1,

$$F = \frac{(n-2)[h(\overline{y}_{t-} - \overline{y}_t)^2 + k(\overline{y}_{t+} - \overline{y}_t)^2]}{h\hat{\sigma}^2_{t-} + k\hat{\sigma}^2_{t+}} = \frac{(n-2)(n\hat{\sigma}^2_t - h\hat{\sigma}^2_{t-} - k\hat{\sigma}^2_{t+})}{h\hat{\sigma}^2_{t-} + k\hat{\sigma}^2_{t+}} \ , \tag{11}$$

where $\hat{\sigma}^2_t$, $\hat{\sigma}^2_{t-}$ and $\hat{\sigma}^2_{t+}$ are the empirical variances calculated from the whole window and from the left and the right-hand windows, respectively, with denominators n, h and k.

The empirical variance in the previous formulae can be replaced by any of the robust scale estimators MAD, LSH, Sn and Qn introduced in Section 2.1. By these means, we expect to achieve more robust ANOVA-type rules, for shift detection, in particular, and for the comparison of several groups in general. Appropriate critical values for any sample size can be derived by simulation, assuming the noise to possess a known distribution, such as the Gaussian.

2.3 Robustified rank tests

Another approach to shift detection is via tests based on linear rank statistics. Prominent amongst these are the Wilcoxon test and the median test (Bovik, Huang and Munson, 1986). Let $y_{t(1)} \leq \ldots \leq y_{t(n)}$ denote the ordered observations in the full window located at time t and let $r_{t,1-h}, \ldots, r_{t,k}$ be the resulting ranks of $y_{t-h+1}, \ldots, y_{t+k}$, which are the positions of these elements in the ordered sequence. For shift detection, a linear rank statistic of the most recent k observations,

$$L_t = \sum_{j=1}^{k} a(r_{t,j}) \ , \tag{12}$$

can be used, where $a(1), \ldots, a(n)$ are given scores. Under H_0 the distribution of L_t is the same for all symmetric noise distributions, i.e. it is distribution-free. In case of bindings, we assign the average rank to identical measurements.

The Wilcoxon test uses the scores $a(i) = i$, $i = 1, \ldots, n$, so that $L_t = \sum_{j=1}^{k} r_{t,j}$. For the median test, we set $a(i) = 1$, $i = \lfloor n/2 \rfloor + 1, \ldots, n$, and $a(i) = 0$ otherwise, so that L_t corresponds to the number of values in $y_{t+1}, \ldots, y_{t+k}$ larger than the overall median $\tilde{y}_t$ from both windows; and it takes values between zero and k.

Fried and Gather (2007) exploit the suggestion of Bovik et al. (1986) to apply a linear rank test after subtraction of a threshold $\tilde{\delta}$ from $y_{t+1}, \ldots, y_{t+k}$

to detect only large shifts. Since the chosen $\tilde{\delta}$ should be large compared to the noise standard deviation σ_t, Fried and Gather choose $\tilde{\delta} = \tilde{\delta}_t$ as a fixed multiple $d\hat{\sigma}_{t-}$ of a robust estimate of σ_t, thereby allowing for a time-varying variability. To prevent a few outliers from unduly influencing the test decision, the critical values for L_t are chosen as large as possible under the restriction that we require a shift to be detected if the largest or smallest $\lfloor (k+1)/2 \rfloor$ observations are in the right-hand window. This makes it possible to identify a level shift even in the presence of $\lfloor (k-1)/2 \rfloor$ large outliers. Choosing $h = k = 5$ e.g. allows to distinguish pairs of outliers from level shifts when using the critical values $1+2+8+9+10 = 30$ and 3 for the Wilcoxon and the median test, respectively, obtained by summing the $\lfloor (k+1)/2 \rfloor$ largest and the $k - \lfloor (k+1)/2 \rfloor$ smallest scores. A small false detection rate of e.g. 0.1% in case of Gaussian noise and a constant level can be achieved by preliminary subtraction or addition of a sufficiently large multiple $\tilde{\delta}_t = d\hat{\sigma}_{t-}$ from $y_{t+1}, \ldots, y_{t+k}$ when testing for an upward or downward shift, respectively. Suitable values of d are determined in simulations.

Fried and Gather (2006) find that, in the presence of outliers the ordinary linear rank tests are outperformed by the robustified versions. They also find that Wilcoxon scores have higher power than median scores. From the robust scale estimators introduced in Section 2.1, the Sn and the Qn statistics yield the highest powers in case of small and large window widths, respectively. The tests employing the Qn are better at distinguishing between outlier patches and shifts than tests employing one of the other robust scale estimators.

3 Comparisons

In the following section, we compare the basic attributes of the detection rules described above. After employing analytic means to investigate the resistances to outliers, a simulation study is performed for comparing the performance of the statistics in small samples. The appropriate choice of the widths h and k, and therefore of $n = h + k$, depends on the circumstances in which a filtering procedure is applied. To mitigate the misleading effects of patches of outliers and of outliers that occur in the vicinity of a level shift, we should choose large values for h and k. However, upper limits are imposed on the length of the windows by the limitation of periods in which the level μ_t can be regarded as virtually constant. Also, the effect of increasing the value of k is to increase delay between the occurrence of a level shift and its detection. We concentrate on circumstances where the windows are small and of equal lengths, $h = k$. This corresponds to the assumption that μ_t is virtually constant only in short windows. In the simulations, the perpetual noise u_t is Gaussian $N(0, 1)$ if not stated otherwise.

3.1 Test resistances

Median filters are popular on account of their robustness in circumstances where a large proportion of the sample is affected by outliers. Breakdown points are analytic measures of the robustness of an estimator. The finite-sample replacement breakdown point of the median is 0.5 asymptotically, which means that modifying less than half of the data cannot drive the estimate beyond all limits.

Resistances are a related concept for measuring the robustness of tests (Ylvisaker, 1977). Let $\mathbf{y} = (y_1, \ldots, y_n)$ be the vector of all observations included in the test, and let ϕ be the decision function of the test with $\phi(\mathbf{y}) = 1$ and $\phi(\mathbf{y}) = 0$ meaning rejection and non-rejection of the null hypothesis, respectively. The idea of the resistance to rejection ϵ_R is to measure the minimal fraction of contaminated observations which can force the test to reject the null hypothesis regardless of the other, 'clean' data. Denote by $U_m(\mathbf{y})$ a neighbourhood of the (clean) data vector $\mathbf{y}$ consisting of all contaminated data vectors $\mathbf{z} = (z_1, \ldots, z_n)$ with $z_i \neq y_i$ for at most $0 \leq m \leq n$ positions. The resistances to rejection and to acceptance can be defined to be, respectively,

$$\begin{aligned}
\epsilon_R &= \frac{1}{n} \min\{m \geq 0 : \inf_{\mathbf{y} \in \mathbb{R}^n} \sup_{\mathbf{z} \in U_m(\mathbf{y})} \phi(\mathbf{z}) = 1\} \ , \\
\epsilon_A &= \frac{1}{n} \min\{m \geq 0 : \sup_{\mathbf{y} \in \mathbb{R}^n} \inf_{\mathbf{z} \in U_m(\mathbf{y})} \phi(\mathbf{z}) = 0\} \ .
\end{aligned}$$

The interpretation of the resistance to acceptance ϵ_A is analogous to that of ϵ_R: irrespective of what the clean data $\mathbf{y}$ are, we can always find a way to avoid rejection of the null hypothesis by replacing $\epsilon_R \cdot n$ of the elements of $\mathbf{y}$. Note that we differ from Ylvisaker in allowing changes to occur at arbitrary positions, since this is more appropriate to the structured data considered here, whereas it makes no difference for unstructured data.

It is appropriate to consider resistances, since outliers should neither prevent detection nor cause false detection of level shifts. The resistance to acceptance of the ordinary two-sample t-test is $1/n$. Irrespective of the data, changing one observation can always reduce the difference of the averages to zero; that is to say, within the context of a t-test, a single outlier can mask a shift of any size. The resistance to rejection of the two-sample t-test is more difficult to calculate. To increase the squared test statistic by an arbitrary amount, so that the p-value goes to zero and thus becomes smaller than any significance level, requires the modification of at least $\min\{h, k\}$ out of the total of n observations. However, the effect of fewer modifications can be large enough to make the test statistic exceed certain significance levels.

We assume $k \leq h$ from now on. Two-sample t-tests based on α-trimmed means and α-winsorized variances resist outliers better than ordinary two-sample t-tests. The resistance to acceptance becomes $(\ell+1)/n$ with $\ell = k\lfloor \alpha k \rfloor$. If the observations in each window are close to each other in value and differ largely from those in the other window, then at least ℓ observations in the right-hand window need to be moved to change the value of the test statistic, which can be reduced to zero. Trimming can reduce the resistance to rejection, since modifying $k - \ell$ observations in the right-hand window can always drive the p-value to zero.

The resistance to acceptance of the robustified rank tests is at least $\min\{\lfloor (k+1)/2 \rfloor, h\epsilon_-^*\}/n$ if we tune the tests in the manner described above (Fried and Gather, 2006). Here, ϵ_-^* is the explosion breakdown point of the scale estimator $\hat{\sigma}_-$ derived from the left-hand window, with explosion meaning breakdown to infinity. The resistance to rejection is at least $\lfloor (k+1)/2 \rfloor /n$.

The resistance to acceptance of a median comparison is at least $\min\{\lfloor (k+1)/2 \rfloor /n, \epsilon^*\}$, where ϵ^* is the explosion breakdown point of the scale estimator. The explanation is the same as for the two-sample t-tests based on α-trimmed means, taking into account that a second possible cause of acceptance is that the scale estimate used for standardization becomes very large. To drive the p-value to zero, we need to modify at least $\min\{\lfloor (k+1)/2 \rfloor, n\epsilon_*\}$ of the values in both windows. Here, ϵ_* is the implosion breakdown point of the scale estimator for data with all values being different. Whereas explosion means breakdown to infinity, implosion means breakdown to zero. This occurs if the scale estimate for a sample can be made arbitrarily small (i.e. close to zero) by replacing some of the observations. If none of the observations is repeated, then we need to modify at least $\lfloor (k+1)/2 \rfloor /n$ values in the right-hand window to make the test statistic arbitrarily large, or to reduce the scale estimate to zero. For the median comparisons based on a joint scale estimate obtained from both windows, both resistances equal $\lfloor (k+1)/2 \rfloor /n$ if we use a highly robust scale estimator. When estimating the variability in the windows separately, however, the resistance to acceptance can be determined by the explosion breakdown point of the scale estimator as it is $k\epsilon^*_-/n$ (only one of the two estimates needs to become too large).

Driving the p-value of ANOVA tests to unity needs a fraction of $\min\{k\epsilon^*_+/n, \epsilon_*\}$ modifications. For the p-value to go to zero the fraction is at least $\min\{(h\epsilon_{*-} + k\epsilon_{*+})/n, \epsilon^*\}$, where ϵ_{*-} and ϵ_{*+} is the implosion breakdown point of the scale estimator from the left and the right-hand windows, respectively.

The number of outliers which a test for shift detection can resist without becoming unreliable depends not only on the window widths h and k, but also on the significance level. Here, we tune all tests to obtain a significance level of 0.1% under Gaussian noise, so that we expect to detect a level shift incorrectly only once in 1000 observations. We set both window widths to the same value $h = k$ for simplicity. This also provides some protection against unequal variances in the two windows (see Staudte and Sheather, 1990).

3.2 Power under different types of noise

Now we compare the power of the tests for different heights of the shift δ. 10000 windows were generated, in each case, for $\delta = 0, 0.5, 1, \ldots, 10$, and the power is derived as the percentage of cases in which a shift is detected. We present the results for the ordinary and for the 30%-trimmed two-sample t-test, the median comparison with joint scale estimation by the MAD or Qn according to (4), the median comparison with separate scale estimation by Sn or Qn (5), the ANOVA test employing Sn or Qn (11), and the robustified Wilcoxon tests (12).

Fig. 2 shows the results for $h = k = 9$ and for a standard Gaussian white noise. As was expected, the ordinary two-sample t-test is the most powerful method of shift detection followed by the median comparisons with joint scale estimators and then by those with separate scales. All of these tests are more powerful than the 30%-trimmed t-test. The ANOVA and the robustified Wilcoxon tests are less powerful. The ANOVA based on Qn misses even huge shifts, because Qn is not sensitive to shifts.

Identical measurements due to rounding, for example, pose a problem for robust scale estimators. A simple remedy is 'wobbling' by a preliminary addition of random noise with the same magnitude as the rounding error. To analyse such effects, we generated shifts of different heights within unit Gaussian noise, as before, and we rounded all observations to the nearest 0.5. In the absence of a shift, more than 95% of the probability is concentrated on the nine values $-2, -1.5, \ldots, 1.5, 2$ then. We added uniform $U(-0.25, 0.25)$ noise to all values to recover the full range. The results were almost identical to those presented above.

Fig. 2 compares the power of the rules under noise generated from a t-distribution with three degrees of freedom, which possesses heavier tails than the Gaussian distribution. All procedures loose some power compared to the Gaussian case. The ordering of the rules remains almost the same except for the ordinary t-test, which looses its superiority and is outperformed by the median comparison with joint MAD. The median tests with separate Sn and with joint Qn are very close to each other and, again, they outperform the 30%-trimmed t-test.

We performed the same experiments as before for other window widths. As expected, the power of all the methods increases with increasing windows, while the differences between the robust approaches are less that they would be with shorter windows. Generally, the orderings of the methods with respect to their power were very similar to those reported before both for Gaussian and for t_3 noise. Therefore, we report only the differences. For $h = k = 7$, the power of the 30%-trimmed t-test dropped down below that of the robustified Wilcoxon test with Sn. For widths larger than $h = k = 9$, the robustified Wilcoxon tests gained power relatively to the other methods.

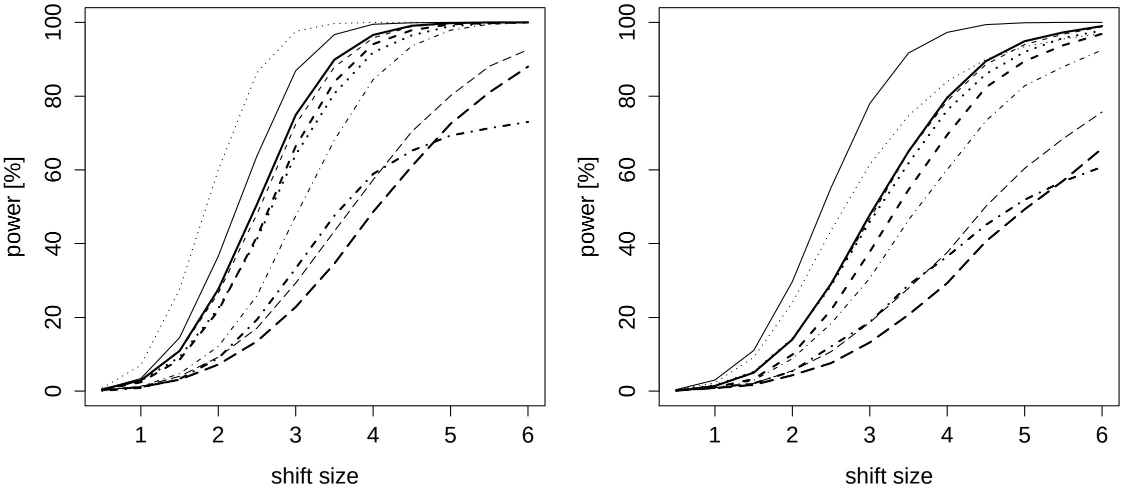

Figure 2: Power for different shift heights, Gaussian (left) and t_3-noise (right): t-test (dotted), 30%-trimmed t-test (bold dotted), median comparison with joint MAD (solid) or Qn (bold solid), with separate Sn (dashed) or Qn (bold dashed), ANOVA on Sn (dash-dot) or Qn (bold dash-dot), Wilcoxon with Sn (wide-dashed) or Qn (bold wide-dashed).

3.3 The case of a single outlier

Next we check the sensitivity of the methods in respect of an outlier of various sizes $s = 1, 2, \ldots, 20$ added to one of the observations.

Fig. 3 shows the error probability of a Type I error caused by an outlier in the right-hand window for $h = k = 9$, estimated from the fraction of cases in which a shift was detected within 50000 simulation runs. The size of the ordinary t-test decays to zero since the test statistic tends to 1 as the outlier size goes to infinity. The median comparisons with separate scale estimates or with joint Qn show a slightly decreasing size, while the ANOVA tests, the 30%-trimmed t-test and the median comparison with joint MAD are almost unaffected. In case of the robustified Wilcoxon tests, we observe a small increase of the error rate, while their size seems to be slightly reduced when the outlier is in the left-hand window (not shown here; note the asymmetry of the robustified Wilcoxon tests due to estimating the scale from the left-hand window and subtracting a multiple of it from the right-hand observations).

We also investigated the power in case of a positive shift of height $\delta = 8\sigma$ and a single positive outlier of size $s = 1, \ldots, 20$ in the left-hand window, or a negative outlier in the right-hand window. Fig. 3 shows the powers obtained from 10000 simulations runs each. The power of the two-sample t-test approaches zero as the outlier size increases, while the 30%-trimmed t-test and the median comparisons are not affected at all. ANOVA tests are affected if the outlier is of the same size as the shift. Robustified Wilcoxon tests are unaffected if the outlier is in the right-hand window, and slightly affected if it is in the left-hand window, with the effect remaining constant as the outlier size exceeds 4σ.

The same results were obtained for windows of width six or seven, while, for $h = k = 15$, only the ordinary two-sample t-test was affected, with both its size and its power going to zero with increasing outlier size.

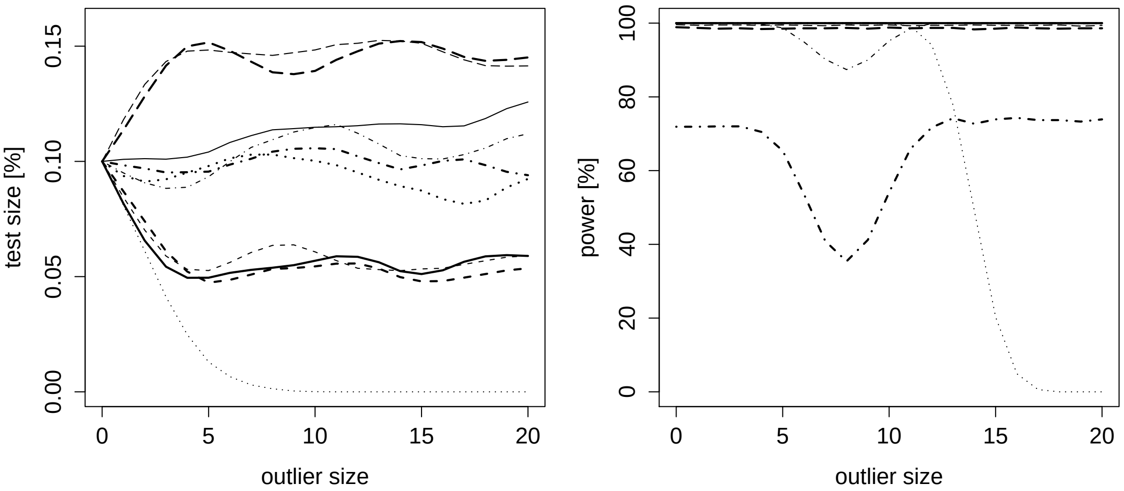

Figure 3: Test size (left) and power for a shift of size 8σ (right) in case of a single outlier of increasing size in the right window: t-test (dotted), 30%-trimmed t-test (bold dotted), median comparison with joint MAD (solid) or Qn (bold solid), with separate Sn (dashed) or Qn (bold dashed), ANOVA with Sn (dash-dot) or Qn (bold dash-dot), Wilcoxon with Sn (wide-dashed) or Qn (bold wide-dashed).

3.4 The case of multiple outliers

For an examination of the rules in case of multiple outliers, we replaced an increasing number of observations in one window by outliers of the same size s. Fig. 4 shows the percentage cases in which a shift was detected within 10000 simulations runs each in case of $s = 8$ and $h = k = 9$. We found analogous results for the widths $h = k = 7$ and $h = k = 15$.

The t-tests only detect a shift with high probability if at least seven out of nine observations are shifted. This is not desirable since a shift is likely to be missed, even when two thirds of the observations deviate from the previous level, a situation pointing more at a shift in combination with a few outliers than at a constant signal overlaid by many outliers. Similar remarks apply to ANOVA tests. Median comparisons with joint scale estimation show a consistent behaviour, since they indicate a shift if more than half of the observations in one window deviate from those in the other window. The median comparison with separate Qn also performs consistently, while six deviating observations are needed when using a separate Sn. All robustified Wilcoxon tests are consistent if the outliers are in the right-hand window, but only the one with Qn performs rather consistently if the outliers are in the left-hand window. We obtained similar findings for the size $s = 12$, with the robustified Wilcoxon test based on Qn giving much better results. Note that the problems of the t-tests and the median comparisons with separate scales were expected given the resistances reported in Section 3.1.

3.5 The case of increasing variance

A phenomenon which should not be confused with a level shift is an increase of the variability. Therefore, we analyse the test sizes when the standard deviation σ_{t+j} in the right hand window becomes $100\%, 120\%, \ldots, 400\%$ of that in the left hand window. A decrease of σ_{t+j} is less interesting, since it reduces the test size.

Fig. 5 depicts the results for $h = k = 9$. All methods indicate a shift more often than in the homoskedastic case. The size of the median comparison with

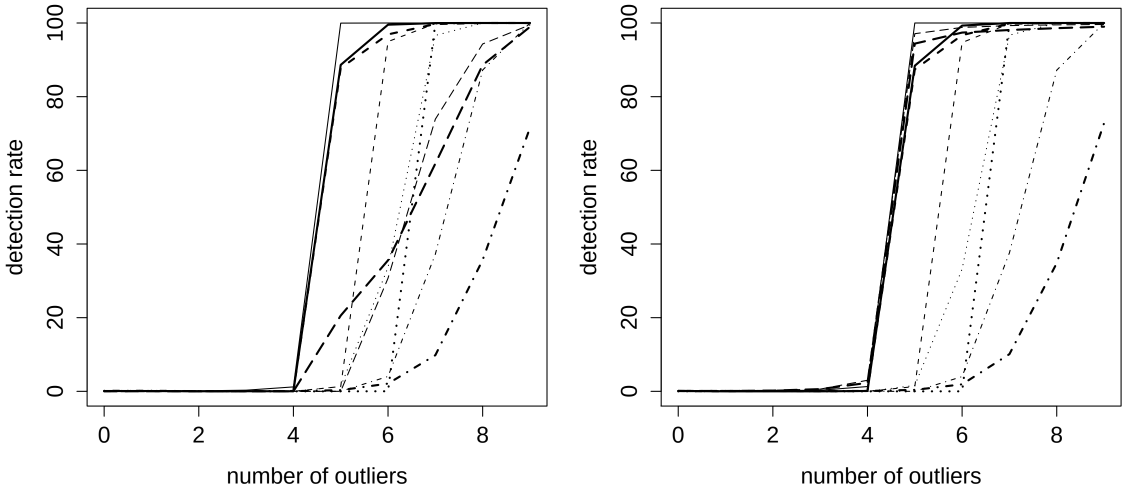

Figure 4: Detection rate in case of an increasing number of outliers of size 8σ in the left or the right window: t-test (dotted), 30%-trimmed t-test (bold dotted), median comparison with joint MAD (solid) or Qn (bold solid), with separate Sn (dashed) or Qn (bold dashed), ANOVA with Sn (dash-dot) or Qn (bold dash-dot), Wilcoxon with Sn (wide-dashed) or Qn (bold wide-dashed).

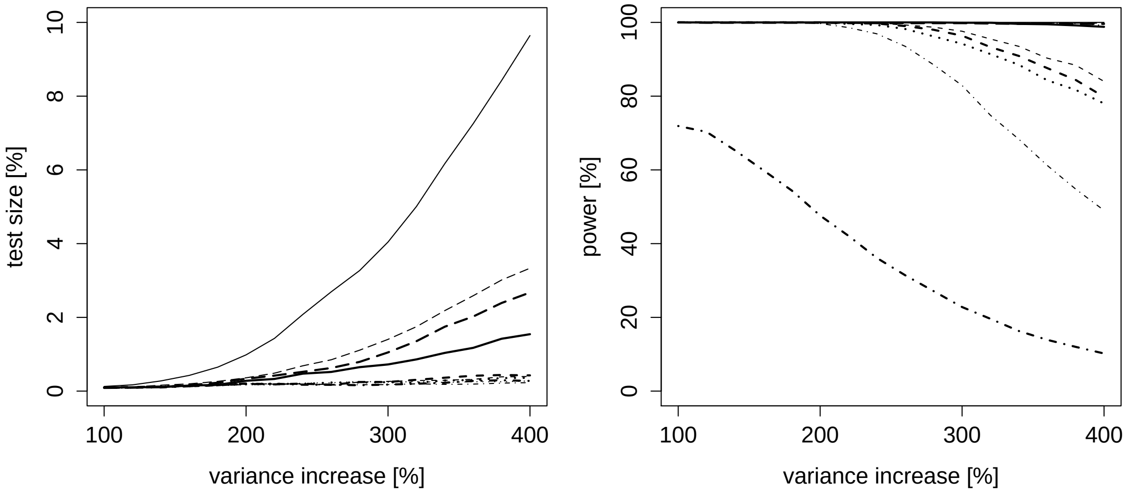

Figure 5: Test size (left) and power for a shift of size 6σ (right) in case of an increase of σ to $x\%$ in the right window: t-test (dotted), 30%-trimmed t-test (bold dotted), median comparison with joint MAD (solid) or Qn (bold solid), with separate Sn (dashed) or Qn (bold dashed), ANOVA with Sn (dash-dot) or Qn (bold dash-dot), Wilcoxon with Sn (wide-dashed) or Qn (bold wide-dashed).

joint MAD goes up to 10%, while for Qn it stays below 2%. As expected, separate scale estimation protects against different variabilities: the increase is only up to about 0.4%, as it is in the case of the ordinary t-test. For ANOVA tests it is even smaller. For robustified Wilcoxon tests we observe an increase to over 2%, which is the larger the more powerful the rule is according to Section 3.2.

We have also investigated the power in case of a shift of size 6σ and a simultaneous increase of σ_{t+j} to $100\%, 120\%, 140\%, \ldots, 400\%$, see also Fig. 5. Those methods which almost keep their size loose a lot of power, namely the ordinary two-sample t-test, the median comparisons with separate scale estimates and particularly the ANOVA tests. Median comparisons with joint scales and robustified Wilcoxon tests keep their power but not their size, as we have seen before. Almost identical results were obtained for other window widths.

3.6 The case of autocorrelations

In many applications measurements are autocorrelated. Fried and Gather (2005) find it better not to modify median filters in the presence of positive autocorrelations, so we continue to use the standard filtering procedures. To investigate the performance of the detection rules in these circumstances, the observational noise was generated from an AR(1) model, $u_t = \phi u_{t-1} + \epsilon_t$, where the innovations ϵ_t constitute a Gaussian white-noise sequence with mean zero and variance $\sigma^2 = 1$. In that case, the noise variance is $\sigma_u^2 = \sigma^2/(1-\phi^2)$ where $\phi = -0.9, \ldots, 0.9$ is the lag-one correlation.

Fig. 6 shows the results for $h = k = 9$. Generally, the increase of the test size seems to be directly related to the power under Gaussian noise, which is reported in Section 3.2. More powerful methods show a larger increase of the size in the case of positive correlations, particularly the t-tests and the median comparisons, while robustified Wilcoxon tests are least affected. The test sizes of the robustified Wilcoxon and the median comparisons using a separate scale remain small as ϕ increases until $\phi = 0.5$ is reached.

An investigation of the power in the case of a shift of height $6\sigma_u$ and for different values of ϕ shows that there is a substantial loss of power under negative correlations, whereas the loss is small for positive ϕ. The ordering of the methods differs little from the ordering in the case of independent errors, $\phi = 0$. Amongst the median comparisons with separate scales, Qn has the highest power for negative ϕ, and the power of the Wilcoxon test with Qn increases in case of large negative ϕ. When investigating the powers in case of a shift of increasing height for fixed positive $\phi = 0.6$, we found the same ordering of the methods as in the case of independent disturbances. Again, we obtained very similar results for other window widths.

4 Application

Finally, we have analysed a time series of length $N = 500$, see Fig. 7. The underlying signal μ_t resembles the blocks function (Donoho and Johnstone, 1994), which is a benchmark example for edge-preserving smoothing. The

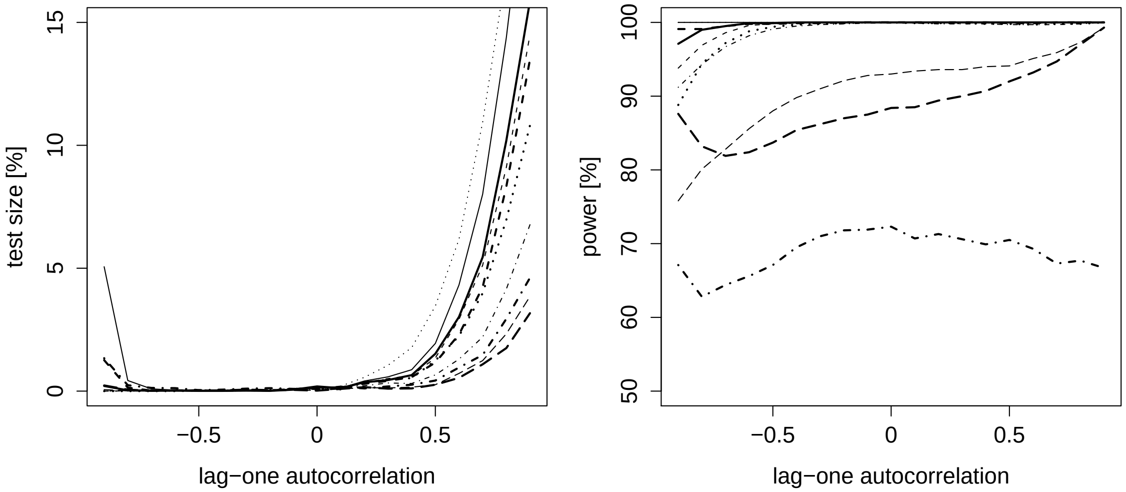

Figure 6: Test size (left) and power for a shift of size $6\sigma_u$ in case of autocorrelations of different size: t-test (dotted), 30%-trimmed t-test (bold dotted), median comparison with joint MAD (solid) or Qn (bold solid), with separate Sn (dashed) or Qn (bold dashed), ANOVA with Sn (dash-dot) or Qn (bold dash-dot), Wilcoxon with Sn (wide-dashed) or Qn (bold wide-dashed).

signal was overlaid by independent Gaussian noise with a time-varying, signal-dependent standard deviation of $\sigma_t = 1 + |\mu_t|/20$. We replaced 40 observations by outliers, adding the same constant $s = 12$ to the observations. Of these, ten were isolated outliers. Another ten outliers came in five pairs, a further twelve came in four triplets, and the remaining eight came in two clusters of four.

A running median with window width $n = 19$ was used for filtering. For detection of a shift at a time point $t \in \mathbb{N}$ with a small delay and for the avoidance of unnecessary alarms we compared the subwindows $y_{t-9}, \ldots, y_{t-1}$ and $y_{t+1}, \ldots, y_{t+9}$. The insertion of a gap between the windows improves the detection of shifts consisting of subsequent steps. Note that we need windows of widths of at least nine points to resist patches of four outliers.

Detection of a shift allows us to take an appropriate action. We apply the method of Fried and Gather (2007) for estimating the time of the level shift: if a shift is detected at time point t but not at $t-1$, the likely time of the shift is immediately before the first $t+j$, $j > 0$, for which y_{t+j} is closer to the median

$\tilde{\mu}_{t+}$ of $y_{t+1}, \ldots, y_{t+9}$ than to the median $\tilde{\mu}_{t-}$ of $y_{t-9}, \ldots, y_{t-1}$. Instead of the median of the full window, we then use the median of the observations in the left window as filter output until time point $t + j - 1$. From $t + j$ on, we use the median of the right window, and return to the median of the full window at time $t + j + 5$.

Fig. 7 also shows various filter outputs. The ordinary running median smoothes the signal edges to some extent. The filter applying the trimmed t-test shows some additional spikes e.g. at time t=112, and it also smoothes the shifts, since these are often detected quite late. Wilcoxon and most ANOVA tests (not shown here) perform better, but they do not overcome the problems completely. Only the median comparison with joint MAD detects the shift at $t = 380$ in a timely fashion. Overall, the median comparisons with joint scale estimate perform best, with the MAD-based version confirming its good power.

5 Conclusions

We have investigated rules for detecting shifts in the presence of outliers. From the results of our experiments we have derived some recommendations on how to proceed when choosing windows as short as those treated here: We have shown that the new ANOVA-type procedures are outperformed by suitably designed median comparisons. In case of homoskedastic noise, the median comparison with a joint MAD scale is recommended if high robustness and detection power are crucial. If the variability varies over time, then joint Qn or separate Sn estimation might be preferred. Unless the windows are very short, we prefer Qn over Sn since its increasing efficiency leads to higher power. Further experiments show that the powers of robustified Wilcoxon tests increase strongly with the width of the window used for the scale estimation, becoming comparable to those of the median comparisons.

These results have been derived for white noise. They remain valid in case of small to moderate autocorrelations. High positive autocorrelations lead to monotonic patterns, which can be confused with level shifts. The corresponding increase of the size of the detection rules can be reduced by estimating the autocorrelations robustly and adapting the thresholds for detection.

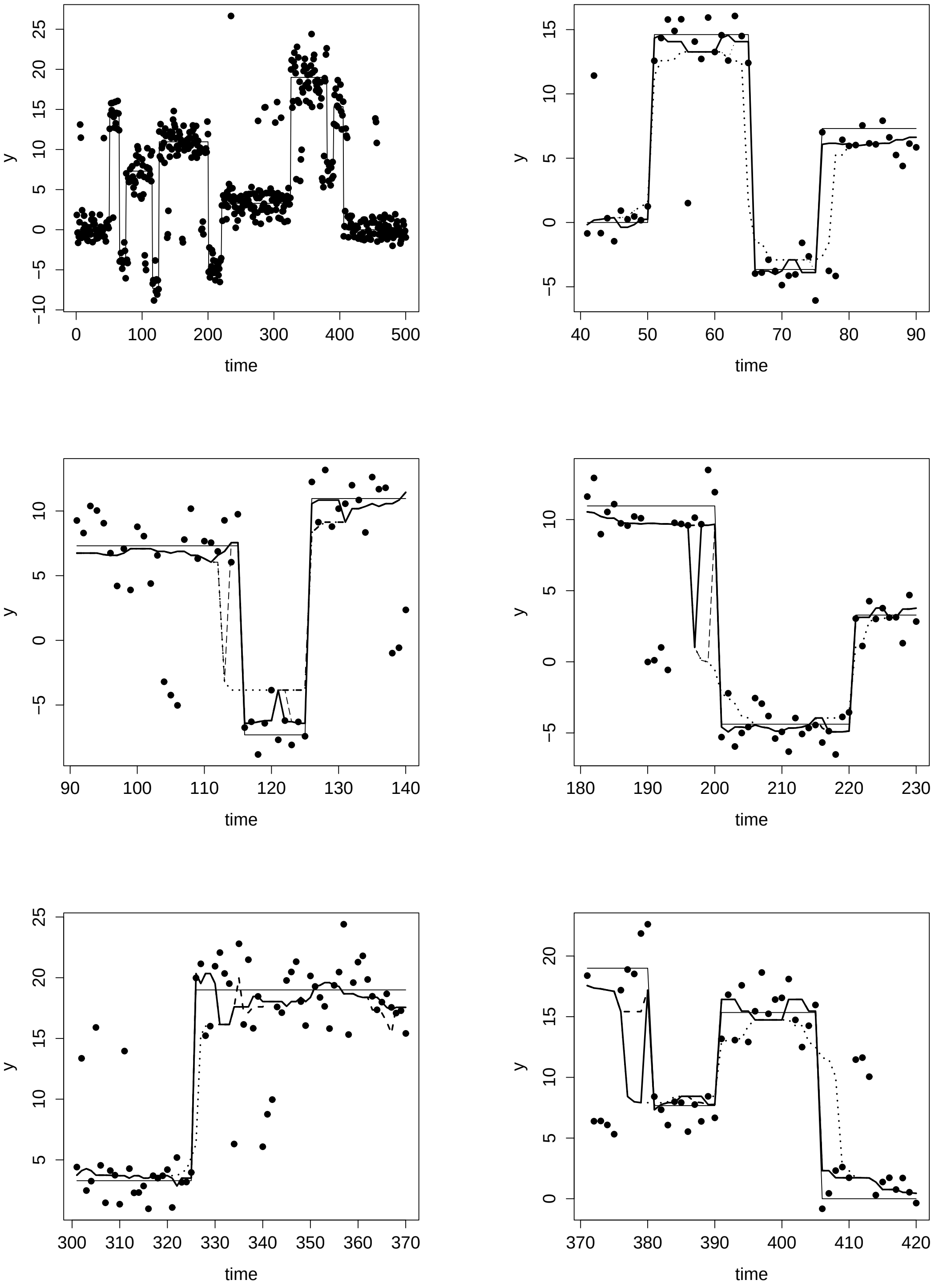

Figure 7: Time series generated from the blocks function (top left) and different time periods with extracted signals: running median (bold dotted) and running median with trimmed t-test (dotted), median comparison with joint Qn (bold solid) or MAD (bold dashed, often coincides with Qn), and Wilcoxon test with Sn (wide-dashed).

We also tried other rules that have been suggested in the literature. t-tests based on ranks (Conover and Iman, 1981) turned out to be almost as powerful as ordinary t-tests, whereas the 20%-trimmed t-test was only slightly worse in this respect. However, the 20%-trimmed t-test protects at most against a single outlier in case of the window widths considered here, and t-test on ranks had little robustness against outliers at all. Among tests based on local variabilities such as those based on quasi-ranges (Restrepo and Bovik, 1988, Sun and Venetsanopoulos, 1988, Kundu and Wu, 1989, Lee and Tantaratana, 1990, Sun, Gabbouj and Neuvo, 1994), only the empirical variance gave good power, but for the price of a strong increase of the test size already in case of a single outlier or a change of the variance. Similar problems were observed with linear hybrid edge detectors (Neuvo, Heinonen and Defee, 1987).

The tests investigated here can be combined with robust regression methods applied recursively to the incoming data. This allows to robustify recursive least-squares techniques for studying the stability of regression relationships over time (Brown, Durbin and Evans, 1975). First experiments show the suitability of Wilcoxon-type tests for the detection of abrupt shifts within trends.

Acknowledgements

The financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures") is gratefully acknowledged.

References

Bovik, A.C., Huang, T.S., Munson, D.C. Jr., 1986. Nonparametric tests for edge detection in noise. Pattern Recognition 19, 209–219.

Bovik, A.C., Munson, D.C. Jr., 1986. Edge detection using median comparisons. Computer Vision, Graphics, and Image Processing 33, 377–389.

Brown, R.L., Durbin, J., Evans, J.M., 1975. Techniques for testing the constancy of regression relationships over time. J. Roy. Statist. Soc. B 37, 149-163.

Cipra, T., 1992. Robust exponential smoothing. J. Forecasting 11, 57–69.

Conover, W.J., Iman, R.L., 1981. Rank transformations as a bridge between parametric and nonparametric statistics. Amer. Statist. 35, 124–128.

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

Fried, R., 2007. Robust location estimation under dependence. J. Statist. Comput. Simul. 77, 131–147.

Fried, R., Gather, U., 2005. Robust trend estimation under AR(1) disturbances. Austrian J. of Statistics 34, 139–151.

Fried, R., Gather, U., 2007. On rank tests for shift detection in time series. Computational Statistics and Data Analysis, to appear.

Gather, U., Fried, R., 2003. Robust estimation of scale for local linear temporal trends. Tatra Mountains Mathematical Publications 26, 87–101.

Grübel, R., 1988. The length of the shorth. Ann. Statist. 16, 619–628.

Hou, Z., Koh, T.S., 2003. Robust edge detection. Pattern Recognition 36, 2083–2091.

Hwang, H., Haddad, R.A., 1994. Multilevel nonlinear filters for edge detection and noise suppression. IEEE Trans. Signal Processing 42, 249–258.

Imhoff, M., Bauer, M., Gather, U., Fried, R., 2002. Pattern detection in intensive care monitoring time series with autoregressive models: influence of the model order. Biometrical J. 44, 746–761.

Kundu, A., Wu, W.-R., 1989. Double-window Hodges-Lehman (D) filter and hybrid D-median filter for robust image smoothing. IEEE Trans. Acoustics, Speech, Signal Process. 37, 1293–1298.

Lee, Y.H., Tantaratana, S., 1990. Decision-based order statistic filters. IEEE Trans. Acoustics, Speech, Signal Process. 38, 406–420.

Neuvo, Y., Heinonen, P., Defee, I., 1987. Linear-median hybrid edge detectors. IEEE Trans. Circuits Systems 34, 1337–1343.

Nieminem, A., Neuvo, Y., Mitra, U., 1989. Algorithms for real-time trend detection. Signal Processing 18, 1–15.

Restrepo, A., Bovik, A.C., 1988. Adaptive trimmed mean filters for image restoration. IEEE Trans. Acoustic, Speech, Signal Process. 36, 1326–1337.

Rousseuw, P.J., Croux, C., 1993. Alternatives to the median absolute deviation. J. Am. Statist. Assoc. 88, 1273–1283.

Rousseeuw, P.J., Leroy, A.M., 1988. A robust scale estimator based on the shortest half. Statist. Neerlandica 42, 103–116.

Staudte, R.G., Sheather, S.J., 1990. Robust Estimation and Testing. John Wiley & Sons, New York.

Stein, R.A., Fowlow, T.J., 1985. The use of median filters for edge detection in noisy time series. Proceedings of ISCAS 85, 1331–1334.

Stigler, S.M., 1973. Laplace, Fisher, and the discovery of the concept of sufficiency. Biometrika, 60, 439–445.

Sun, T., Gabbouj, M., Neuvo, Y., 1994. Adaptive L-filters with applications in signal and image processing. Signal Processing 38, 331–344.

Sun, X.Z., Venetsanopoulos, A.N., 1988. Adaptive schemes for noise filtering and edge detection by use of local statistics. IEEE Trans. Circuits Systems 35, 57–69.

Tukey, J. W., 1977. Exploratory Data Analysis. Addison-Wesley, Reading, Mass. (preliminary edition 1971).

Ylvisaker, D., 1977. Test resistance. J. Am. Statist. Assoc. 72, 551–556.

Yuen, K.K., 1974. The two-sample trimmed t for unequal population variances. Biometrika 61, 165–170.

Zeileis, A., 2005. A unified approach to structural change tests based on ML scores, F statistics, and OLS residuals. Econometric Reviews 24, 445–466.