SFB
823

Control charts for the mean based on robust two-sample tests

Sermad Abbas, Roland Fried

Discussion Paper

Nr. 21/2016

SFB
823

Control charts for the mean based on robust two-sample tests

S. Abbas * and R. Fried

Faculty of Statistics, TU Dortmund University, Vogelpothsweg 87, 44221 Dortmund, Germany

We propose and investigate robust control charts for the detection of sudden shifts in sequences of very noisy observations with a naturally slowly varying mean. They sequentially apply local two-sample tests for the location problem. Thus, no previous knowledge about the in-control behaviour is necessary.

We identify critical values for the tests to achieve a desired in-control average run length (ARL_0) with extensive simulations. Control charts based on nonparametric tests or a randomization principle provide a satisfactory run length behaviour for different error distributions. They possess a nearly distribution-free ARL_0 and are fast in detecting present signal jumps in a time series.

In our simulations and exemplary real-world applications from biosignal analysis, a test based on the two-sample Hodges-Lehmann estimator leads to very promising results regarding distribution independence, robustness and detection speed.

Keywords: biosignal analysis; change-point detection; robust control charts; time series; two-sample tests; monitoring

AMS Subject Classification: 62G10; 62G35; 62M10; 62L10; 62P10

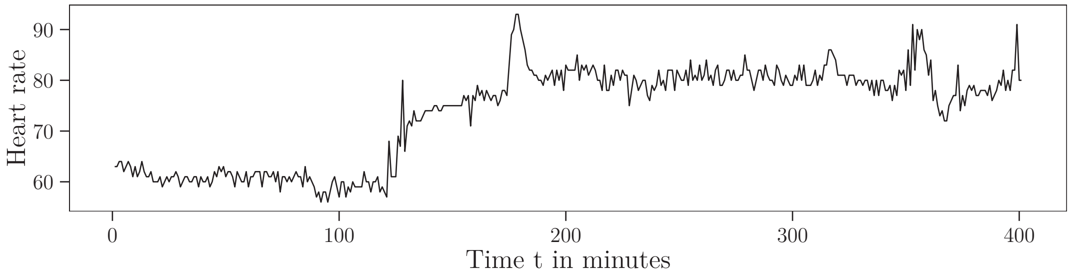
1. Introduction

The detection of sudden changes in the signal underlying a time series is an important task in biosignal analysis. We concentrate on situations where the data are observed subsequently in equidistant time intervals and tolerate slow variations in the mean.

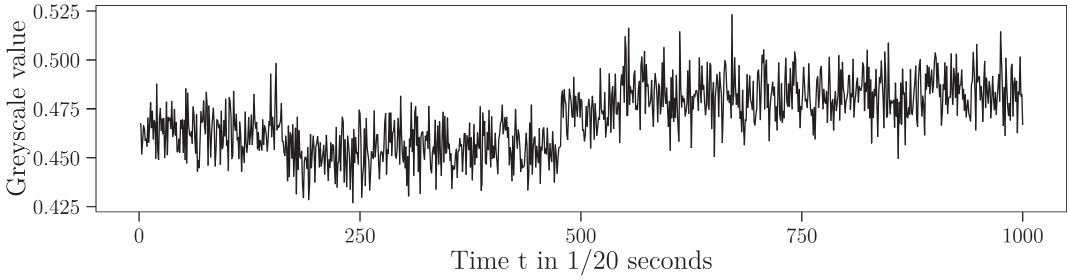
For example, in intensive care vital parameters, e.g. the heart rate of a person, are monitored. Abrupt changes in the signal can indicate clinically relevant events. Figure 1(a) shows the heart rate of a patient. A large level shift begins at time $t = 121$. In addition, some patches of very large values can be seen, e.g. at $t = 178$. The challenging task is to distinguish between relevant and irrelevant changes which are caused, e.g., by measurement artefacts or movements of the patient [1]. An appropriate method should detect the relevant changes quickly while ignoring unimportant ones. Moreover, the number of false alarms should be as small as possible because of a potential alarm fatigue by the medical staff [2, 3].

Another example is the plasmon assisted microscopy of nanosize objects (PAMONO). The PAMONO biosensor is used to check a sample fluid for the existence of specific objects with a size on the nanoscale, e.g. viruses. If a virus adheres on the sensor surface, a permanent bright spot surrounded by a dark circle appears in a sequence of greyscale images taken from the surface. To analyse the data, a time series of greyscale values is extracted for each pixel coordinate. If the coordinate is part of a virus adhesion, it has

*Corresponding author. Email: abbas@statistik.tu-dortmund.de



(a) Exemplary time series from the online heart-rate monitoring in intensive care.



(b) Exemplary time series from the plasmon assisted microscopy of nanosize objects for retrospective virus detection. Relevant level shifts occur at $t = 164, 478, 539$.

Figure 1.: Examples for the detection of abrupt signal changes in time series.

a sudden positive or negative jump depending on whether the coordinate belongs to the bright spot or the dark circle. More details on the sensor can be found in [4] or [5]. Figure 1(b) shows a time series from a PAMONO data set. It contains three shifts induced by virus adhesions at times $t = 164, 478, 539$. Between these change points, the signal is approximately constant. There are no outliers and the time series behaves better than in the intensive-care example.

Obvious methods to monitor such data are control charts. They typically work in two phases. Phase I is used to set up the control limits. In Phase II the actual monitoring is performed by comparing a sequentially calculated control statistic to these limits. If the limits are violated, an action is taken to adjust the process [7]. In [8] the necessity for robust methods is pointed out. To get appropriate control limits, the process has to be in control, meaning that it is not influenced by any kind of non-random disturbances. Often this cannot be assured so that the Phase-II results might be negatively influenced. This is one reason why classical charts like the Shewhart, CUSUM (Cumulative Sum) or EWMA (Exponentially Weighted Moving Average) control charts are not appropriate in some applications. They depend on certain reference values or historical in-control data to estimate them. For example, in the intensive care context a common reference value or in-control phase does not exist [1]. In the PAMONO application, an adjustment after a level shift is unnecessary. Moreover, multiple level shifts can occur in a time series. Thus, a reference value would have to be updated after each shift to detect the next one. This is difficult because the magnitude of a shift is unknown.

Control charts for nonlinear time series based on the CUSUM and EWMA approaches have been investigated in, e.g., [9, 10] but also depend on reference values or historical data.

Our aim is to construct robust methods which work without prior knowledge of the data. Thus, we can overcome the problems with Phase I. To take natural variations in the signal into account, the procedures should adapt to the level of the time series signal. Moreover, we prefer nonparametric approaches because we do not want to make any assumption on the distributional origin of the data. In recent years, nonparametric procedures have become more and more prominent in the control chart context. Overviews are given in [11] and [12]. Theses procedures have in common that they depend on fixed reference values or samples.

We use two-sample tests in a moving time window to detect sudden changes. The window is split in a reference and a test window which are compared by the test statistic. Thus, it is unnecessary to specify a reference value. We examine control charts based on the two-sample t-test, the Wilcoxon rank-sum test and the median test. Furthermore, we consider tests based on the difference of the sample medians and the one-sample Hodges-Lehmann estimators as well as the two-sample Hodges-Lehmann estimator [13, 14]. They serve as robust alternatives to the t-test. In [14], only classical test properties, i.e. the size and power of the aforementioned tests, are investigated in different scenarios. Here, we evaluate the control chart behaviour of detection procedures based on sequential two-sample testing. We use simulations to investigate their run length, which is the duration between two successive alarms. In the in-control situation, when there is no shift in the signal, the run length should be large. When the process is out of control, i.e. after a shift, the run length should be small. We consider the average run length (ARL) and the median run length (MRL). The methods are compared in situations with and without level shifts under several types of distributions.

In [15] a control chart based on the one-sample Hodges-Lehmann estimator is proposed as an alternative to the Shewhart control chart. In [16] it is stated that the chart is not able to achieve a desired in-control ARL and cannot be regarded as distribution free. Two further control charts using the Hodges-Lehmann estimator are presented in the article, which are both able to maintain a desired in-control ARL under normality. The first one compares the Hodges-Lehmann estimate for a sample to control limits based on the average of the empirical variance of subgroups in historical data. The second chart uses the control limits from [15] but the control statistic is now a multiple of the Hodges-Lehmann estimator. However, again, both charts are not distribution free and rely on prior knowledge. Furthermore, they differ essentially from our approach.

Our results indicate that the moving-window approach is quite promising. It is possible to construct distribution-free and nearly distribution-free procedures which work well even in situations with slow trends. In addition, our results suggest that control procedures based on robust test statistics inherit their robustness against outliers.

The outline of the paper is as follows: In Section 2 the basic model is introduced. Section 3 presents the two-sample tests we will study. In Section 4 the results of the simulation studies to asses the performances of the procedures are described. In Section 5 we illustrate the practical suitability of our approach by applying the procedures to the intensive care and PAMONO time series. Section 6 summarizes and discusses the results.

2. Model

Let $(Y_t : t \in \mathbb{N})$ be a time series which is decomposed using the additive components model

$$Y_t = \mu_t + \varepsilon_t + \eta_t,\ t \in \mathbb{N}. \tag{1}$$

Here, $(\mu_t : t \in \mathbb{N})$ is the time-dependent unknown underlying signal. It is assumed to be smooth and to follow a slow, possibly nonmonotonic trend, with only a few abrupt jumps. The independent and identically distributed random variables $(\varepsilon_t : t \in \mathbb{N})$ describe additive random noise with expectation $\mathrm{E}(\varepsilon_t) = 0$ and constant variance $\mathrm{Var}(\varepsilon_t) = \sigma^2 > 0$. The process $(\eta_t : t \in \mathbb{N})$ represents an outlier-generating mechanism which is usually zero but sporadically leads to large absolute values. All random variables ε_t and η_t are supposed to be independent.

We use a moving time window of width $n = h + k$, $h, k \in \mathbb{N}$, to detect abrupt shifts in the signal of a time series generated from model (1). The window at time t,

$$\boldsymbol{Y}_t = (Y_{t-h+1}, \ldots, Y_t, Y_{t+1}, \ldots, Y_{t+k})',\ t = h, h+1, \ldots,$$

is split into two subwindows. They are called the reference window $\boldsymbol{Y}_{t-}$ of width h and the test window $\boldsymbol{Y}_{t+}$ of width k. We rename the random variables in both subwindows so that

$$\boldsymbol{Y}_{t-} = \left(Y_{t,1}^-, \ldots, Y_{t,h}^-\right)' \text{ and } \boldsymbol{Y}_{t+} = \left(Y_{t,1}^+, \ldots, Y_{t,k}^+\right)',$$

where

$$Y_{t,i}^- = Y_{t-h+i},\ i = 1, \ldots, h, \text{ and } Y_{t,j}^+ = Y_{t+j},\ j = 1, \ldots, k.$$

We assume that the signal is constant within both subwindows. In order to check whether a sudden signal jump occurs between the times t and $t+1$, we will use two-sample tests for the location problem to compare the test window with the reference window as in [17]. If there are no outliers in the time series, i.e. $\eta_t = 0$ for all $t \in \mathbb{N}$, the expected values in the subwindows are given by

$$\mathrm{E}\left(Y_{t,i}^-\right) = \mu_{t-h+i} = \mu_{t-},\ i = 1, \ldots, h, \text{ and } \mathrm{E}\left(Y_{t,j}^+\right) = \mu_{t+j} = \mu_{t+},\ j = 1, \ldots, k,$$

where μ_{t-} and μ_{t+} are constants with $\mu_{t+} = \mu_{t-} + \Delta_t$. Here, $\Delta_t \in \mathbb{R}$ is the unknown jump magnitude of the signal between t and $t+1$. The variances are

$$\mathrm{Var}\left(Y_{t,i}^-\right) = \sigma^2,\ i = 1, \ldots, h, \text{ and } \mathrm{Var}\left(Y_{t,j}^+\right) = \sigma^2,\ j = 1, \ldots, k.$$

Thus, the underlying distributions in both subwindows differ at most in location, so that

$$Y_{t,1}^+, \ldots, Y_{t,h}^+ \overset{\text{i.i.d.}}{\sim} F \text{ and } Y_{t,1}^-, \ldots, Y_{t,k}^- \overset{\text{i.i.d.}}{\sim} G,$$

where $F,\ G : \mathbb{R} \to [0,1]$ are the distribution functions of the underlying continuous distributions with $G(x) = F(x - \Delta_t)$ for all $x \in \mathbb{R}$.

Under the null hypothesis, $H_{0,t} : \Delta_t = 0$, there is no jump between t and $t+1$. We call a rejected null hypothesis an alarm. An incorrectly rejected null hypothesis (type I error) will be referred to as a false alarm.

According to [18], a benefit of the moving-window approach is that there is no need to fit a global parametric model to the data. Thus, no assumptions on the global behaviour of the signal have to be made.

The subwindow widths h and k have to be chosen under consideration of the application. Large values help to reduce the influence of outliers in the area of a signal jump. Subwindows which are too small can cause outliers to have a large effect on the test decision so that they could be mistaken for a signal change or mask existent changes. If the subwindow widths are too large, the assumption of a locally constant signal may not be justified. In addition, the time between the occurrence and the detection of a jump gets larger. Moreover, large window widths can lead to an easier confusion between a trend and a true location shift, i.e. the rejection of the null hypothesis is induced by the trend [18].

3. Methods

In this section, procedures for the detection of sudden changes in the signal underlying a time series will be introduced. We apply two-sample tests for the location problem in moving time windows to avoid the need for reference values. Furthermore, we are able to adapt to the temporal development in the time series. In Subsections 3.1 and 3.2 we present several test statistics which are the basis of our control procedures. A detailed description of how to achieve a test decision follows in Subsection 3.3. Subsection 3.4 then deals with some criteria to analyse the run length behaviour of the methods.

The time index t will be dropped in the following because we focus on a single window.

3.1. *Test statistics based on estimating the location difference*

Our test statistics are based on comparing the reference and the test window by estimating the location difference and standardizing it with a suitable scale estimator.

A popular example for this principle is the two-sample t-test. The difference of the sample means

$$\hat{\Delta}^{(0)}(\boldsymbol{Y}) = \overline{Y}_+ - \overline{Y}_-, \text{ where } \overline{Y}_+ = \frac{1}{k}\sum_{j=1}^{k} Y_j^+ \text{ and } \overline{Y}_- = \frac{1}{h}\sum_{i=1}^{h} Y_i^-,$$

is standardized by the pooled empirical standard deviation

$$\hat{S}^{(0)}(\boldsymbol{Y}) = \sqrt{\frac{1}{n-2}\left(\sum_{i=1}^{h}\left(Y_i^- - \overline{Y}_-\right)^2 + \sum_{j=1}^{k}\left(Y_j^+ - \overline{Y}_+\right)^2\right)}.$$

This leads us to the test statistic

$$T^{(t)}(\boldsymbol{Y}) = \sqrt{\frac{h \cdot k}{n}} \cdot \frac{\hat{\Delta}^{(0)}(\boldsymbol{Y})}{\hat{S}^{(0)}(\boldsymbol{Y})}.$$

If the random variables in both subwindows follow a normal distribution with equal but unknown variances, $T^{(t)}$ follows a t-distribution with $n-2$ degrees of freedom under the null hypothesis. Because of the central limit theorem, the type I error will also be controlled well, if the distributions are not normal, but the subwindow widths are sufficiently large [19, p. 240]. However, the t-test is not robust against outliers since a few can cause a substantial loss in power or an exceedance of the significance level [17].

It is possible to construct robust tests by replacing the sample means and the pooled standard deviation by robust alternatives [14]. An obvious choice to estimate the location difference robustly is the difference of the sample medians,

$$\hat{\Delta}^{(1)}(\boldsymbol{Y}) = \widetilde{Y}_{+} - \widetilde{Y}_{-},$$

where $\widetilde{Y}_{+} = \operatorname{med}\left\{Y_1^{+}, \ldots, Y_k^{+}\right\}$ and $\widetilde{Y}_{-} = \operatorname{med}\left\{Y_1^{-}, \ldots, Y_h^{-}\right\}$. We consider two robust scale estimators given by

$$\hat{S}^{(1)}(\boldsymbol{Y}) = \operatorname{med}\left\{|Y_1^{-} - \widetilde{Y}_{-}|, \ldots, |Y_h^{-} - \widetilde{Y}_{-}|, \ldots, |Y_1^{+} - \widetilde{Y}_{+}|, \ldots, |Y_k^{+} - \widetilde{Y}_{+}|\right\}$$

and the sum of the median absolute deviations for both subwindows

$$\hat{S}^{(2)}(\boldsymbol{Y}) = \operatorname{MAD}(\boldsymbol{Y}_{+}) + \operatorname{MAD}(\boldsymbol{Y}_{-}),$$

where

$$\begin{aligned}\operatorname{MAD}(\boldsymbol{Y}_{+}) &= 1.4826 \cdot \operatorname{med}\left\{|Y_1^{+} - \widetilde{Y}_{+}|, \ldots, |Y_k^{+} - \widetilde{Y}_{+}|\right\} \text{ and}\\ \operatorname{MAD}(\boldsymbol{Y}_{-}) &= 1.4826 \cdot \operatorname{med}\left\{|Y_1^{-} - \widetilde{Y}_{-}|, \ldots, |Y_h^{-} - \widetilde{Y}_{-}|\right\}.\end{aligned}$$

The factor 1.4286 is used for correction to achieve an asymptotically unbiased estimation of the standard deviation under the normal distribution [20, p. 33]. The resulting test statistics are

$$T^{(\mathrm{MD1})}(\boldsymbol{Y}) = \frac{\hat{\Delta}^{(1)}(\boldsymbol{Y})}{\hat{S}^{(1)}(\boldsymbol{Y})} \text{ and } T^{(\mathrm{MD2})}(\boldsymbol{Y}) = \frac{\hat{\Delta}^{(1)}(\boldsymbol{Y})}{\hat{S}^{(2)}(\boldsymbol{Y})}.$$

The tests are called MD1- and MD2-test in the following.

An often mentioned drawback of the sample median is its low efficiency in comparison to the sample mean under the normal distribution [21]. We therefore consider some estimators which lead to a compromise between robustness and efficiency.

The one-sample and the two-sample Hodges-Lehmann estimators have an asymptotic relative efficiency of $\frac{3}{\pi} \approx 0.95$ under the normal distribution compared to the sample mean and are considerably more robust [13].

The one-sample Hodges-Lehmann estimator applied to both subwindows with

$$\widehat{Y}_{+} = \operatorname{med}\left\{\frac{Y_i^{+} + Y_j^{+}}{2} : 1 \le i < j \le k\right\} \text{ and } \widehat{Y}_{-} = \operatorname{med}\left\{\frac{Y_i^{-} + Y_j^{-}}{2} : 1 \le i < j \le h\right\}$$

leads to the location-difference estimator

$$\hat{\Delta}^{(2)}(\boldsymbol{Y}) = \widehat{Y}_{+} - \widehat{Y}_{-}.$$

The two-sample Hodges-Lehmann estimator calculates the pairwise differences between the observations in the subwindows:

$$\hat{\Delta}^{(3)}(\boldsymbol{Y}) = \operatorname{med}\left\{Y_i^+ - Y_j^- :\ i = 1, \ldots, k,\ j = 1, \ldots, h\right\}.$$

In [14] two scale estimators, which estimate the variability within the subwindows, are suggested. The estimator

$$\hat{S}^{(3)}(\boldsymbol{Y}) = \operatorname{med}\left\{|Y_i^- - Y_j^-| :\ 1 \le i < j \le h,\ |Y_i^+ - Y_j^+| :\ 1 \le i < j \le k\right\}$$

calculates the median of the absolute pairwise differences within the subwindows. It is related to the two-sample Hodges-Lehmann estimator.
With

$$\hat{S}^{(4)}(\boldsymbol{Y}) = \operatorname{med}\left\{|Z_i - Z_j| :\ 1 \le i < j \le n\right\},$$

where

$$(Z_1, \ldots, Z_n)' = \left(Y_1^- - \widetilde{Y}_-, \ldots, Y_h^- - \widetilde{Y}_-, Y_1^+ - \widetilde{Y}_+, \ldots, Y_k^+ - \widetilde{Y}_+\right)',$$

we calculate the median of the absolute differences within the whole window. The random variables in each subwindow are corrected by the corresponding sample median. The resulting test statistics are

$$T^{(\mathrm{HL11})}(\boldsymbol{Y}) = \frac{\hat{\Delta}^{(2)}(\boldsymbol{Y})}{\hat{S}^{(3)}(\boldsymbol{Y})},\ T^{(\mathrm{HL21})}(\boldsymbol{Y}) = \frac{\hat{\Delta}^{(3)}(\boldsymbol{Y})}{\hat{S}^{(3)}(\boldsymbol{Y})}$$
$$T^{(\mathrm{HL12})}(\boldsymbol{Y}) = \frac{\hat{\Delta}^{(2)}(\boldsymbol{Y})}{\hat{S}^{(4)}(\boldsymbol{Y})},\ T^{(\mathrm{HL22})}(\boldsymbol{Y}) = \frac{\hat{\Delta}^{(3)}(\boldsymbol{Y})}{\hat{S}^{(4)}(\boldsymbol{Y})}$$

and the corresponding tests will be called HL11-, HL21-, HL12- and HL22-test.

3.2. *Test statistics based on linear rank statistics*

Let $R_1^-, \ldots, R_h^-, R_1^+, \ldots, R_k^+$ be the ranks of $Y_1^-, \ldots, Y_h^-, Y_1^+, \ldots, Y_k^+$ in the joint sample. The underlying distributions are assumed to be continuous and thus, the probability of assigning the same rank to two observations is zero. Nevertheless, in applications the observed values are typically rounded and two of them could be equal. We will assign the ranks randomly in such cases.
We consider two different linear rank tests: The two-sample median test and the two-sample Wilcoxon rank-sum test.

The test statistic of the two-sample median test counts the number of observations in the test window which are larger than the median of the whole window. It is given by

$$T^{(\mathrm{M})}(\boldsymbol{Y}) = \sum_{j=1}^{k} \mathbb{1}\left(R_j^+ > \frac{n+1}{2}\right),$$

where $\mathbb{1}(A)$ is the indicator function with condition A. Under the null hypothesis, the test statistic follows a hypergeometric distribution.

Another popular nonparametric test is the two-sample Wilcoxon rank-sum test. The test statistic is the sum of the ranks in the test window $\boldsymbol{Y}_+$, i.e.

$$T^{(\mathrm{W})}(\boldsymbol{Y}) = \sum_{j=1}^{k} R_j^+.$$

The distribution of the test statistic under the null hypothesis can be derived by a permutation principle [22, p. 113].

The test statistics of the median and the Wilcoxon test follow a discrete distribution. We use randomization to achieve exact significance levels [23, p. 24].

3.3. *Derivation of a test decision*

For the t-test the distribution of the test statistic under the null hypothesis is known under normality. The distributions of the rank-test statistics are also known under the null hypothesis. For the MD- and the HL-tests the distribution under the null hypothesis is unknown in finite samples. In [14] it is proposed to use the permutation principle to construct distribution-free tests. In that case, all $B = \binom{n}{k}$ splits of the complete window in two subwindows can be determined. For each split the value of the test statistic is calculated. Let $T_1, \ldots, T_B$ be the test statistics for each permutation and T_{obs} the observed value. The p-value for a two-sided test is

$$p = \frac{\sum_{i=1}^{B} \mathbb{1}\left(|T_i| \geq |T_{\mathrm{obs}}|\right)}{B}. \tag{2}$$

The enumeration of all possible splits of a window is computationally demanding because B increases fast with growing h and k. An alternative approach is to take a random number $b \leq B$ of all possible splits additionally to the observed one, to derive a p-value. Again, the value of the test statistic is calculated for each sample, leading to a randomization distribution. Let $T_1, \ldots, T_{b+1}$ be the test statistics calculated for the selected splits and T_{obs} the observed value. Then, a p-value is derived as

$$\hat{p} = \frac{\sum_{i=1}^{b} \mathbb{1}\left(|T_i| \geq |T_{\mathrm{obs}}|\right) + 1}{b+1}.$$

We select the splits with replacement which is computationally easier because we do not have to check for each split whether it was already drawn [24].

Although the randomization leads to smaller computation times for a single test, the long-term monitoring of a time series requires the sequential application of the test. A computation from scratch for each time window might, especially in high-frequency applications, lead to unacceptable large computing times. An example for this is given in Section 4. Hence, we consider two strategies to calculate a fixed reference distribution which will be used to make all test decisions in the time series.

In the first one, we use the randomization distribution of the observations in the first time window of the time series as the reference distribution. By doing this, we implicitly assume that the distributional class of all following observations is the same so that the reference distribution is still appropriate for later time windows. For the sake of simplicity, the tests which use this procedure will be called randomized tests. This distribution can be recalculated at later points in time if necessary.

Our second approach is to simulate the distribution under the null hypothesis by making a distributional assumption for the observations. We study this approach under the normality assumption. We draw N random samples of size n from the standard normal distribution and split it into two subwindows of size h and k. The test statistic is then calculated for each split. We estimate the p-value by using formula (2) where B is replaced by N. This procedure has the benefit that we do not have to recompute the distribution for each new time series. However, the procedures depend on a distribution. In the remainder of this paper we will refer to them as simulative tests.

3.4. *Selected criteria for the run length analysis*

In this subsection, we assume that only one jump of height $\Delta \in \mathbb{R}$ occurs in the signal. If there is no location shift, i.e. $\Delta = 0$, the process is said to be in control, otherwise it is out of control.

The standard approach to compare statistical tests is to study their size and power. Here, we use the tests to create control procedures. In this context, their quality is typically measured by using the run length R which is the number of observations until the first alarm. It should be large when the process is in control and small when it is out of control. We assume that the out-of-control situation is present from the beginning of the monitoring, i.e. the change and the surveillance start at the same time [25]. We summarize the run length distribution by the popular average run length (ARL), which is the expected duration until the first alarm and depends on the shift height Δ,

$$\mathrm{ARL}(\Delta) = \mathrm{E}_{\Delta}\left(R\right)$$

[26]. For $\Delta = 0$ this is named in-control ARL (ARL_0). If $\Delta \neq 0$, it is the expected duration until the detection of the location shift and called the out-of-control ARL (ARL_1). A commonly used way to compare different control procedures is to fix the ARL_0 to a desired value and evaluate the methods in terms of their ARL_1 [27, p. 153]. This works in analogy to the comparison of statistical tests where the significance level is fixed and several tests are compared with respect to their power.

Although the ARL is an often applied criterion, it is confronted with some criticism. A frequently addressed point is that it is not an appropriate measure to represent the run length distribution because the latter can be very skewed. If the run length distribution is right skewed, the ARL will be larger than the majority of the actually achieved run lengths by a control procedure. Therefore, in [28] it is suggested to calculate the median run length (MRL) instead,

$$\mathrm{MRL}\left(\Delta\right) = \mathrm{med}_{\Delta}\left(R\right).$$

The MRL is easier to interpret. For a fixed MRL-value one knows that one half of the run lengths is smaller and the other half is larger. Analogously to the ARL we use the terms in-control MRL (MRL_0) and out-of-control MRL (MRL_1).

In the simulation study in Section 4, where we analyse the in-control behaviour, we use the ARL while we will focus on the MRL in the out-of-control evaluation. The reasoning behind this is the following: In the in-control situations it is important to trigger false alarms rarely and the ARL contains more information on the amount of false alarms than the MRL. In the out-of-control scenario, the MRL delivers more information on the detection speed because a specific MRL_1 indicates that with probability 0.5 a signal change will be detected within the first MRL_1 time points after its occurrence.

As we use control procedures based on statistical tests, we have to specify an appropriate significance level to fix the ARL_0 or MRL_0. In the case of independent time windows, the number of tests until the first alarm is geometrically distributed with detection probability $\alpha \in (0,1)$. Thus,

$$\mathrm{ARL}_0 = \frac{n}{\alpha} \text{ and } \mathrm{MRL}_0 = n \cdot \left\lceil \frac{\log\left(\frac{1}{2}\right)}{\log\left(1-\alpha\right)} \right\rceil .$$

When a moving time window is used, the test statistics are dependent so that the number of tests until the first alarm does not follow a geometric distribution and the relationship between the ARL_0 (MRL_0) and α becomes more difficult to describe. Therefore, in Subsection 4.1, we use simulations to characterize it.

4. Simulations

We assess the performance of the control procedures described in Section 3 in simulations. In Subsection 4.1 we analyse the in-control behaviour. Subsection 4.2 deals with the out-of-control case. The methods are evaluated by examining their ARL and MRL as described in Subsection 3.4. We use the following error distributions:

- Standard normal distribution ($\mathcal{N}(0,1)$)
- t-distribution with five degrees of freedom (t_5)
- t-distribution with two degrees of freedom (t_2)
- χ^2-distribution with three degrees of freedom (χ^2_3), shifted to have expectation zero.

Transferring the remarks of [29, p. 2] to our situation, a control procedure should work well over a wide range of error distributions. The $\mathcal{N}(0,1)$-distribution is treated as an ideal case and serves as a reference in the following. The t_5- and t_2-distributions are examples of heavy-tailed distributions, while the χ^2_3-distribution represents a skewed distribution. It is expected that a procedure which works well for all distributions considered here, will also deliver good results in less extreme situations.
We concentrate on the subwindow widths $h = k = 10$. For the randomized tests, we use $b = 10000$ random samples. The distributions of the simulative tests are calculated with $N = 50000$ random samples from a $\mathcal{N}(0,1)$-distribution.

The simulations are conducted on the Linux HPC cluster LiDo in Dortmund by using the statistical software R, version 3.1.0 [30]. On each node, we have a 3.00 GHz Intel Xeon E5450 machine with 15 GB RAM. The computations are carried out with the R package BatchExperiments [31]. We evaluate the simulation results on our local machine using R, version 3.2.1 [32]. Graphics are created with the R packages ggplot2 [33] and tikzDevice [34].

To illustrate the gain in computational speed when using a fixed randomized reference distribution instead of calculating a new one for each time window, we compare both approaches exemplarily for the randomized HL11-test. We generate 20000 observations from a $\mathcal{N}(0,1)$-distribution in 100 replications and apply both versions to the time series.

The classical randomization test needs, on average, about 12800 seconds. When using a fixed reference distribution, we only have a mean time of approximately 8 seconds.

4.1. *Analysis of the in-control behaviour*

In this subsection, we analyse the control methods in several in-control situations. Our main interest lies in studying the relationship between the ARL_0 and the significance level α. The goal is to choose an appropriate α to achieve a desired ARL_0. Ideally, the relationship between α and the ARL_0 would be distribution free, i.e. a fixed α leads to the same ARL_0 under each error distribution. However, most of the considered control methods depend on a distributional assumption so that this cannot be expected.
We compare the achieved ARL_0-values under the t_5-, t_2- and χ^2_3-distribution to those of the $\mathcal{N}(0,1)$-distribution. When the ARL_0 is larger than for the $\mathcal{N}(0,1)$-distribution, we call the method conservative for this distribution. If it is smaller, the procedure is called anti-conservative.

In Subsection 4.1.1 we present the simulation results. We specify the relationship between α and the ARL_0 in Subsection 4.1.2. For the sake of simplicity, the control methods based on the two-sample tests will be referred to by the name of the underlying test.

4.1.1. ARL_0-curves under different distributions

We generate 20000 time series of length 20000 for each error distribution. The procedures are applied with $\alpha = 0.005, 0.01, 0.015, \ldots, 0.1$. The ARL_0 is estimated by calculating the arithmetic mean over all run lengths for each procedure and α. Although the probability for a false alarm in finite time is one, the time series may be too short to give an alarm. Missing run lengths are replaced by their lower bound 20001 in such cases. This may lead to an underestimation of the true run length. As the proportion of missing values is smaller than 1% in all cases, the effect will be negligible.

We structure the simulation results by splitting the methods into different groups. The first group consists of the randomized control procedures, the second comprises the simulative methods and the third one covers the t-, the median and the Wilcoxon test. Figures 2 - 4 show the achieved ARL_0 as a function of the significance level α for the different distributions. We cut off the y-axis at 400 to emphasize the differences between the different distributions. For all methods one has to keep the randomness of the simulation in mind. Thus, it is possible that the curves for the different distributions intersect in some cases. Hence, we only describe the general tendency regarding conservatism and anti-conservatism.

Figure 2 shows the ARL_0-curves for the randomized procedures. The t-test is slightly conservative for all distributions. Under the t_5-distribution, the procedures behave similar as under the $\mathcal{N}(0,1)$-distribution. Only for the HL1-tests large deviations to the $\mathcal{N}(0,1)$-curve are visible as these procedures have the tendency to be conservative for $\alpha \geq 0.01$. The HL2- and MD-tests, in contrast, behave nearly distribution free. Generally, the deviation to the $\mathcal{N}(0,1)$-curve gets smaller with increasing α for all considered procedures. The scale estimator $\hat{S}^{(4)}$ used for the Hodges-Lehmann based procedures, seems to result in a somewhat smaller difference to the $\mathcal{N}(0,1)$-curve.
In the investigated situations, the randomization principle can lead to approximately distribution-free procedures, but this is not guaranteed. The HL1-tests seem to have problems in situations where the error distribution is heavy tailed or asymmetric. According to [35] the one-sample Hodges-Lehmann estimator is not recommended in such scenarios. Hence, the influence of a distribution on the estimator also has an impact on

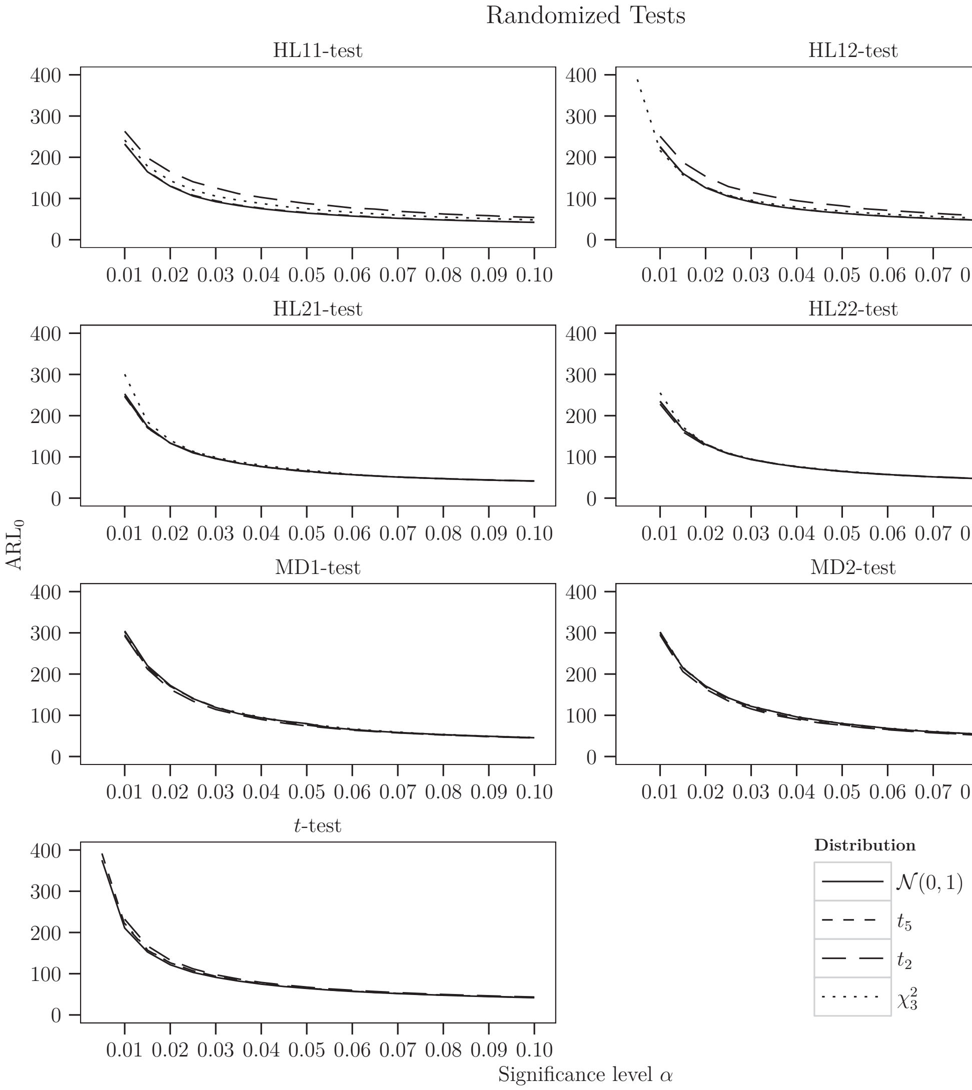


Figure 2.: ARL$_0$ of the randomized control methods as a function of α for different error distributions.

the control procedure.

In Figure 3 the results for the simulative control procedures are presented. All methods are conservative for the heavy-tailed distributions. As expected, the t_2-distribution leads to larger deviations from the $\mathcal{N}(0,1)$-distribution than the t_5-distribution. Except for the HL22- and the t-test, which are slightly conservative, all procedures are anti-conservative under the χ_3^2-distribution. In this respect, the HL22-test seems to be a little more advantageous because the deviations to the $\mathcal{N}(0,1)$-curve are smaller for $\alpha \leq 0.035$ than for the t-test. Considering the heavy-tailed distributions, the HL1-tests lead to good results. Due to the normality assumption, the simulative control procedures are much more prone to the error distribution. Differences between these methods regarding the underlying distribution are mostly visible for small values of α. These are the more interesting cases in applications, as one is often interested in having a large ARL$_0$.

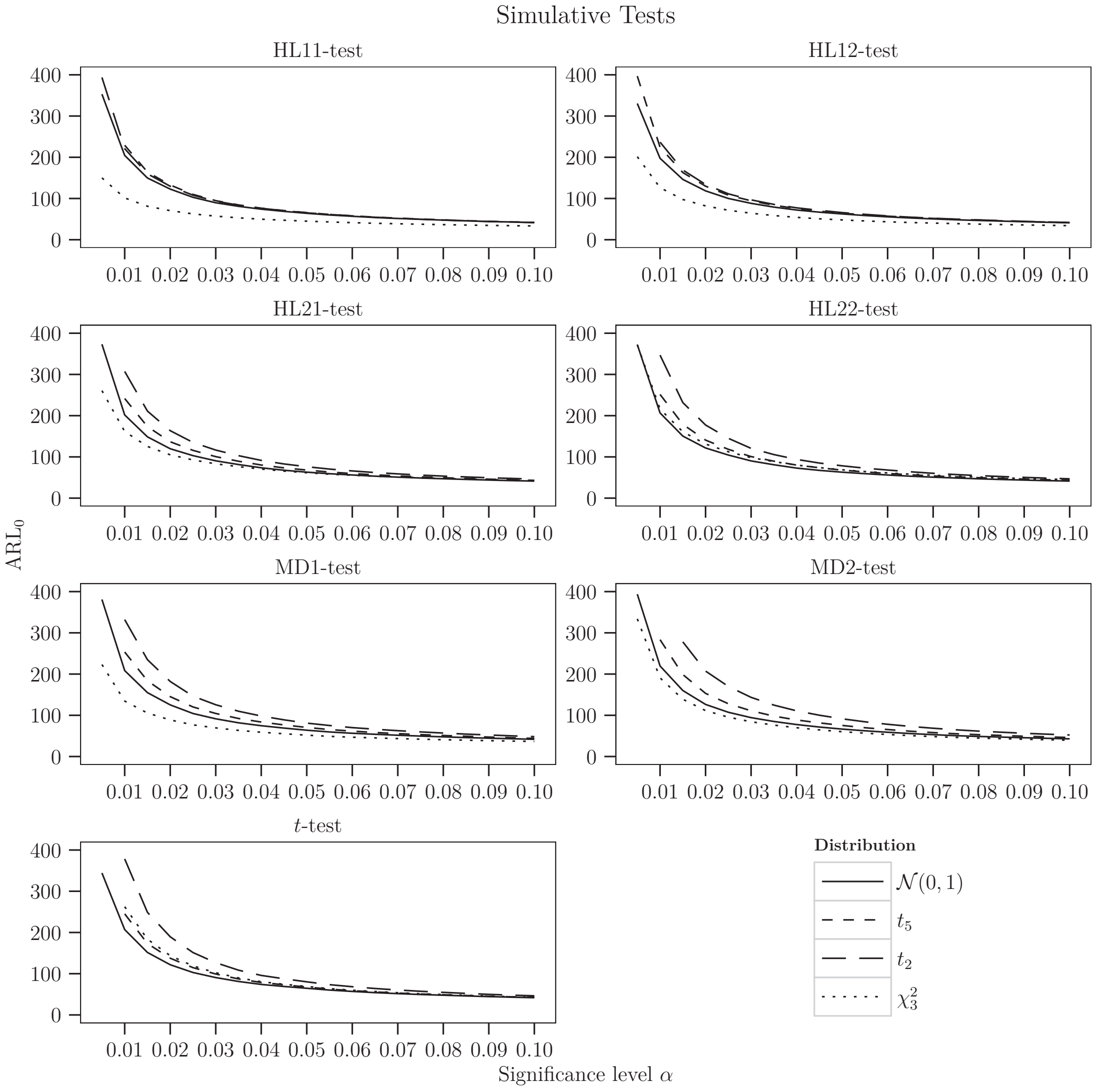


Figure 3.: ARL$_0$ of the simulative control methods as a function of α for different error distributions.

Figure 4 presents the results for the t-, the median and the Wilcoxon test. The Wilcoxon and the median test lead to a distribution-free ARL$_0$. The t-test behaves similar to its simulative version shown in Figure 3.

The MRL$_0$, which is not shown here, has a similar behaviour for the simulative and the classical tests as the ARL$_0$. The main difference is that the MRL$_0$ is generally smaller than the ARL$_0$ because the run length distributions are right-skewed. For the randomized tests we see that now all procedures show a nearly distribution-free behaviour. Only for the HL1-tests some smaller differences occur at the beginning, but these are much less apparent than for the ARL$_0$. The HL1-tests are now clearly anti-conservative for $\alpha \leq 0.02$ for the t_2- and the χ^2_3-distribution. This is also true for the t-test in case of the t_2-distribution.

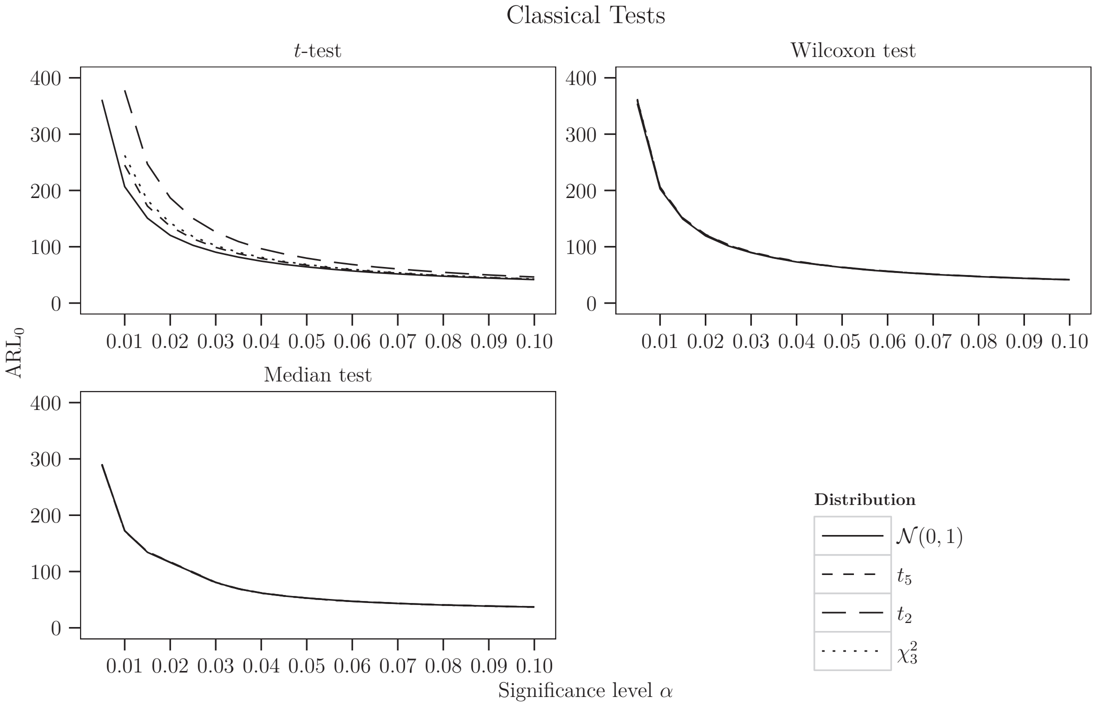


Figure 4.: ARL$_0$ of the control methods based on classical two-sample tests as a function of α for different error distributions.

4.1.2. Relationship between the ARL$_0$ and the significance level

We use our simulation results to specify the functional relationship between α and the ARL$_0$. Figures 2 - 4 suggest the functional form

$$\mathrm{ARL}_0 = \beta_0 \cdot \alpha^{\beta_1}$$

with unknown parameters $\beta_0 > 0$ and $\beta_1 < 0$. This generalises the formula for independent time windows, where $\beta_0 = n$ and $\beta_1 = -1$. Linearisation with the logarithm leads to

$$\log(\mathrm{ARL}_0) = \underbrace{\log(\beta_0)}_{=\gamma_0} + \underbrace{\beta_1}_{=\gamma_1} \cdot \log(\alpha) \Leftrightarrow \alpha = \exp\left(\frac{\log(\mathrm{ARL}_0) - \gamma_0}{\gamma_1}\right). \tag{3}$$

We estimate the parameters γ_0 and γ_1 by ordinary least squares. For the MRL$_0$ the relationship can be described analogously.

4.2. *Analysis of the out-of-control performance*

In this subsection, we compare the control procedures regarding their out-of-control behaviour. We fix the ARL$_0$ in $\{100, 300\}$ and evaluate the performance of the procedures with respect to their MRL$_1$. The necessary values for α are calculated by using relationship (3), repeating the simulations described in Subsection 4.1.1 with smaller spaces between α for a better approximation of the true values. We determine the values under the $\mathcal{N}(0,1)$-distribution. This has to be considered when we interpret the simulation outcome because it will be affected by the conservatism and anti-conservatism of some of our methods.

We generate 10000 time series of length $20000 + h + k + 1$ for each error distribution and insert a permanent location shift starting at time $t = h + k + 1$. The first time window contains the observations at $t = 2, \ldots, h+k+1$ so that only the rightmost observation is shifted and the monitoring begins with the change. The observations at $t = 1, \ldots, h+k$ will be used to calculate the reference distributions for the randomized tests. We use multiples $\Delta = 0.5, 1, 1.5, 2$ of the difference between the 84.13%- and the 50%-quantile of the error distributions as jump heights for a better comparability between the distributions [14]. This difference equals one for the $\mathcal{N}(0,1)$-distribution.

We again replace missing run lengths by 20001. In all considered cases, the number of missing run lengths is smaller than 0.1%.

Table B1 in the appendix shows the smallest MRL_1 for the different error distributions split by the jump heights and both ARL_0-values. We compare the methods by using the relative efficiency which is calculated as the minimal MRL_1 from Table B1 divided by the actually achieved MRL_1 in the different situations. This criterion describes how much worse a procedure is in comparison to the best one. A good method should lead to a large relative efficiency close to one. We calculate worst-case relative efficiencies by computing a method's minimal relative efficiency (MRE) for each jump-height over all distributions. This is in a similar spirit as the Minimax approach for estimators, see [29, p. 60].

First, we concentrate on $\mathrm{ARL}_0 = 100$. The results are shown in the upper row of Figure 5. The randomized tests lead to the best worst-case relative efficiencies for $\Delta = 0.5$ of at least 69%. They are closely followed by the Wilcoxon, the median and the simulative HL1-tests with minimal relative efficiencies of more than 60%. For $\Delta = 1$, the classical and the simulative t-test with values of about 75% perform considerably worse than the other methods which have a minimal relative efficiency of at least 87%. For $\Delta \geq 1.5$ all worst-case relative efficiencies are larger than 90%, so that the differences can be regarded as negligible.

When the t_2-distribution is ignored in the calculation of the minimum, none of the relative efficiencies decreases. Especially for the classical and the simulative t-test, the results gain a comparatively large amount of efficiency for $\Delta = 0.5$. It is now 66% compared to 44% in the situation with all error distributions. The increasing efficiency is not surprising since all procedures are conservative under the t_2-distribution and thus the detection of a jump is delayed. The removal of the χ^2_3-distribution instead of the t_2-distribution has a large effect on the MD2-test where the minimal relative efficiencies are now close to one for all jump heights.

The efficiencies get generally smaller for $\mathrm{ARL}_0 = 300$. Again, the randomized procedures deliver the best results. The removal of the t_2- or the χ^2_3-distribution from the calculation of the minimum has a similar effect as for $\mathrm{ARL}_0 = 100$.

When interpreting the simulation outcome one has to take the in-control results into account. We fixed the ARL_0 to the same value for all procedures, but this does not mean that the MRL_0 will be the same for all methods as well. For example, the MRL_0-values for the randomized MD-tests are about 20 smaller than those for the remaining randomized tests in case of $\mathrm{ARL}_0 = 100$ and approximately 60 for $\mathrm{ARL}_0 = 300$. This means that the probability of an early false alarm is higher which explains why the randomized MD-tests deliver such good results. Except for the randomized MD-tests, only the randomized HL2- and the Wilcoxon and median tests deliver nearly the same MRL_0 under all distributions

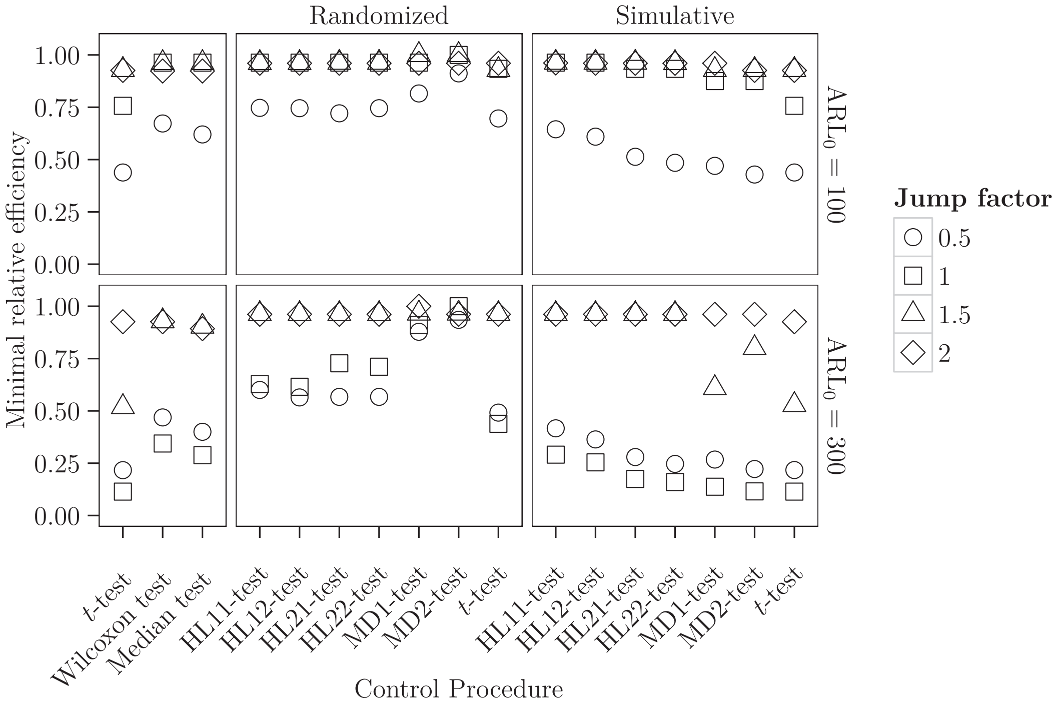


Figure 5.: Minimal relative efficiency based on the MRL for each control method over all considered distributions depending on the jump height.

for our considered ARL_0-values. The HL2-tests have a smaller MRL_0 than the Wilcoxon and the median test. Recommending a specific procedure is therefore difficult. The MD-results seem to be too optimistic because of the small MRL_0. Thus, the method of choice should be one of the nonparametric tests or the randomized HL2-tests. This is confirmed by a comparison of the minimal relative efficiencies regarding the ARL_1. The randomized HL2-tests and the nonparametric tests clearly outperform the randomized MD-tests in this respect (see Figure A1 in the appendix).

5. Applications

In this section, we apply the control procedures based on the *t*-test, the Wilcoxon test and the randomized HL22-test to the time series presented in Section 1. The *t*-test is considered because it is the standard procedure for the two-sample location problem. The Wilcoxon test is its most popular nonparametric competitor and the HL22-test delivered good results in our simulation studies. The significance level will be chosen to achieve $ARL_0 = 300$ under the $\mathcal{N}(0,1)$-distribution. The subwindow widths are $h = k = 10$. Furthermore, we use $N = 10000$ for the HL22-test.

First, we consider the PAMONO time series, where we know the true times of the location shifts. To challenge the procedures, we insert artificial outliers in two different scenarios. Figure 6(a) shows the original time series that does not contain any outlier. The vertical lines in the upper part indicate the alarms given by the three procedures and the true shift times. The methods detect only the first two location shifts. The last shift at $t = 539$ is comparatively small. Considering that the time series consists of 1000 observations, the number of false alarms is fairly small.

Next, we look at two scenarios with one outlier each. We concentrate on the jump at time $t = 478$ because it triggers the highest numbers of alarms. We include either a positive outlier (0.52 instead of 0.44) at time $t = 472$ or a negative one at time $t = 482$ (0.4 instead of 0.47) to mask this shift. The positive outlier (Figure 6(b)) before and the negative one (Figure 6(c)) after the true jump prevent it from being detected by the t-test. The Wilcoxon test now only raises two alarms in the area around the shift while it leads to eight alarms in the outlier-free situation. The HL22-test is less affected as it induces eight alarms in the clean scenario, triggers seven for the positive outlier and six for the negative one. Thus, it is still reliable in both situations.

In the PAMONO context, several further analysis steps are used to decide, whether a structure in the image sequence is caused by a virus. Hence, some false alarms are no major drawback of a method in this scenario. For this time series, all procedures seem to be equally suited.

In the heart-rate time series from intensive care (Figure 7) we cannot be sure about the positions of relevant signal changes. Intuitively, we would regard the large jump at $t = 121$ as relevant. The spikes at $t = 178$ and $t = 355$ could be artefacts but a detection by the procedures is likely because of their duration and magnitude. At $t = 90$ there is a small negative shift. The HL22-test rejections concentrate on the area around the jump at $t = 121$. The t- and the Wilcoxon test lead to a considerably larger number of alarms at the ascending slope after this jump. All methods trigger an alarm at both peaks and for the negative shift. The HL22-test does not lead to rejections at the negative slope starting at $t = 1$ and the peak at $t = 317$, while the t- and the Wilcoxon test do so. Thus, for this time series, the HL22-test delivers promising results as the control procedure only reacts to the larger changes, whereas it ignores smaller fluctuations. This fits well with the objective of reducing the large number of false alarms in such applications [2].

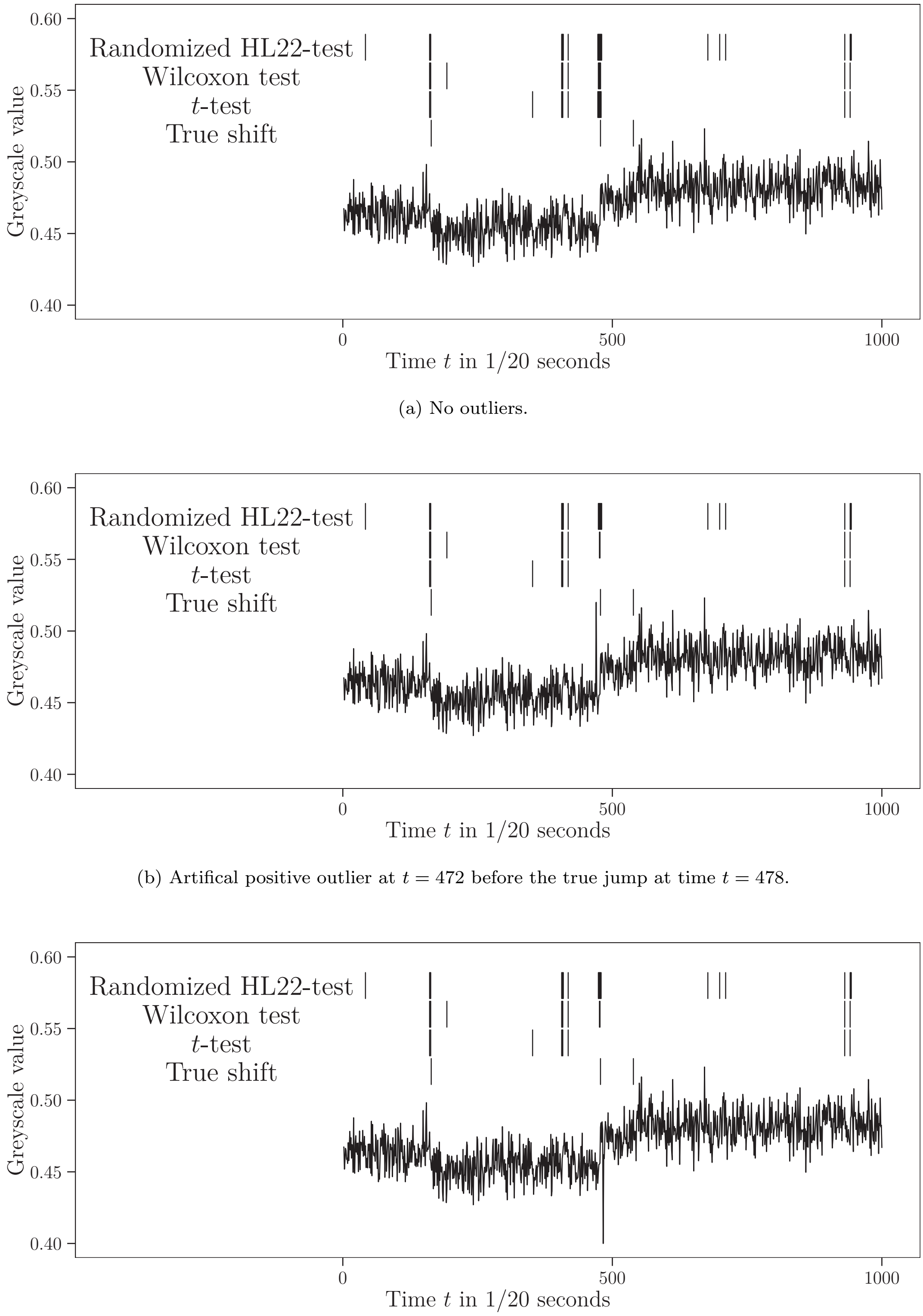
In this context, the question can arise, if the large shift is only detected because of the two positive outliers. Replacing them by smaller values so that they fit into the other observations turns out to have no influence.

6. Conclusion

We study control methods for the detection of abrupt level shifts in time series. The procedures are based on two-sample tests for the location problem in a moving time window. For each time point, we test if it is a change point by splitting the window into two subwindows. They are then compared by the test statistic. In contrast to classical control charts like the Shewhart control chart, the test-based methods do not need a fixed reference value and thus do not depend on historical data or prior knowledge. Furthermore, they can be easily applied in situations with multiple jumps and adapt to the current level of the time series.

We compare procedures based on selected two-sample tests in extensive simulation studies by analysing their run lengths. We examine the average run length (ARL) and the median run length (MRL). We consider different error distributions, i.e. the $\mathcal{N}(0,1)$-distribution, t_5-distribution, t_2-distribution and χ^2_3-distribution in various in- and out-of-control scenarios. Thus, we implicitly deal with outliers in the data.

Control methods based on the Wilcoxon and median test lead to a distribution-free in-control ARL. This has the advantage that we can fix the significance level of a test to achieve a desired in-control ARL under an arbitrary error distribution. Randomized



(a) No outliers.

(b) Artifical positive outlier at $t = 472$ before the true jump at time $t = 478$.

(c) Artifical negative outlier at $t = 482$ after the true jump at time $t = 478$.

Figure 6.: PAMONO time series with two artificial outlier scenarios. The times of the level shifts and the rejection times of selected control procedures are marked by vertical lines.

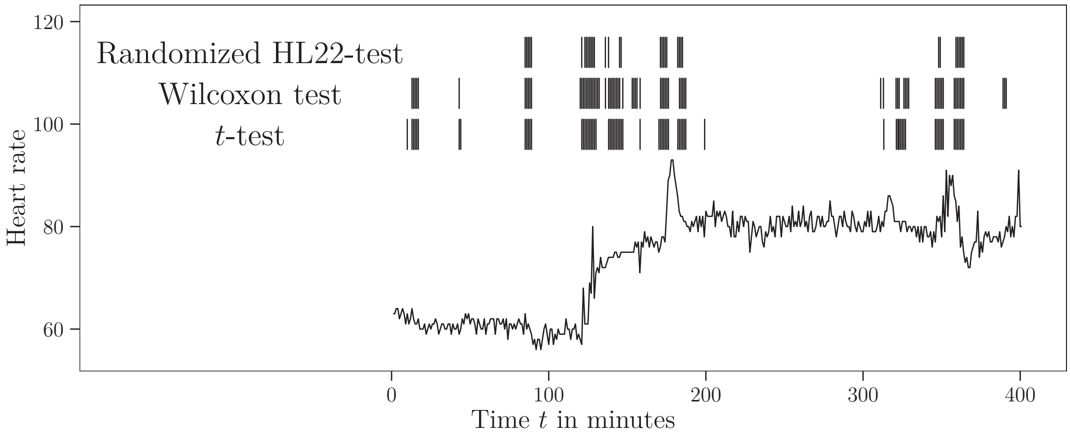


Figure 7.: Heart-rate time series from online monitoring in intensive care. The rejection times of selected procedures are marked by vertical lines.

tests based on the two-sample Hodges-Lehmann (HL2) estimator and the difference of the sample medians (MD) lead to approximately distribution-free in-control ARL-values. Our randomization principle differs from the one classically used because we calculate only one reference distribution based on the first n observations in the time series, where n is the window width. The distribution will be used for all following time windows. This assumes that the distributional structure of the observations does not change over time. If this assumption is dubious, we can update the reference distribution occasionally and thus adapt to the new structure. In the out-of-control situation we use the MRL to compare the methods because of its better interpretability. The randomized HL2-test leads to quite good results in this respect. The Wilcoxon and the median test are only slightly worse for small jump heights.

Two real-world examples indicate that the presented approach is suitable for the detection of sudden changes. Even in the case of small trends, the methods work reliably in the sense that not too many false alarms are triggered. Here, robust methods like the HL2-test delivered better results than non-robust ones like the t-test.

If steeper trends are expected under control, the proposed control charts could be combined with methods for adaptive robust signal extraction as developed in [36]. We then could apply the charts to the residuals of such a procedure.

Acknowledgements

The authors would like to thank the Collaborative Research Center “Providing information by resource-constrained analysis” (SFB 876), Subproject B2, of the German Research Foundation (DFG), for providing the PAMONO data for public use. Thanks also go the ITMC of the TU Dortmund University for providing the HPC cluster LiDo.

This work has been supported by the Collaborative Research Center “Statistical modelling of nonlinear dynamic processes” (SFB 823) of the DFG.

References

[1] Imhoff M, Bauer M, Gather U, Fried R. Pattern detection in intensive care monitoring time series with autoregressive models: Influence of the model order. Biometrical Journal. 2002; 44(6):746–761.

[2] Borowski M, Görges M, Fried R, Such O, Wrede C, Imhoff M. Medical device alarms. Biomedizinische Technik/Biomedical engineering. 2011;56(2):73–83.

[3] Imhoff M, Kuhls S. Alarm algorithms in critical care monitoring. Anesthesia and Analgesia. 2006;102(5):1525–1537.

[4] Zybin A, Kuritsyn YA, Gurevich EL, Temchura VV, Überla K, Niemax K. Real-time detection of single immobilized nanoparticles by surface plasmon resonance imaging. Plasmonics. 2010;5(1):31–35.

[5] Siedhoff D, Libuschewski P, Weichert F, Zybin A, Marwedel P, Müller H. Modellierung und Optimierung eines Biosensors zur Detektion viraler Strukturen. In: Deserno MT, Handels H, Meinzer HP, Tolxdorff T, editors. Bildverarbeitung für die Medizin 2014: Algorithmen - Systeme - Anwendungen: Proceedings des Workshops vom 16. bis 18. März 2014 in Aachen. Berlin: Springer; 2014. p. 108–113.

[6] Collaborative Research Center SFB 876, Project B2. PAMONO Sensor Data 200nm_10Apr13. 2014.

[7] Woodall WH, Montgomery DC. Some current directions in the theory and application of statistical process monitoring. Journal of Quality Technology. 2014;46(1):78–94.

[8] Pan JN, Chen SC. New robust estimators for detecting non-random patterns in multivariate control charts: a simulation approach. Journal of Statistical Computation and Simulation. 2011;81(3):289–300.

[9] Garthoff R, Okhrin I, Schmid W. Statistical surveillance of the mean vector and the covariance matrix of nonlinear time series. Advances in Statistical Analysis. 2014;98(3):225–255.

[10] Garthoff R, Okhrin I, Schmid W. Control charts for multivariate nonlinear time series. REVSTAT. 2015;13(2):131–144.

[11] Chakraborti S, Van der Laan P, Bakir ST. Nonparametric control charts: An overview and some results. Journal of Quality Technology. 2001;33(3):304–315.

[12] Chakraborti S, Human S, Graham MA. Nonparametric (distribution-free) quality control charts. In: Balakrishnan N, editor. Handbook of methods and applications of statistics: Engineering, quality control, and physical sciences. Wiley, New York; 2011. p. 298–329.

[13] Hodges JL, Lehmann EL. Estimates of location based on rank tests. The Annals of Mathematical Statistics. 1963;34(2):598–611.

[14] Fried R, Dehling H. Robust nonparametric tests for the two-sample location problem. Statistical Methods & Applications. 2011;20(4):409–422.

[15] Alloway JA, Raghavachari M. Control chart based on the Hodges-Lehmann estimator. Journal of Quality Technology. 1991;23(4):336–347.

[16] Pappanastos EA, Adams BM. Alternative designs of the Hodges-Lehmann control chart. Journal of Quality Technology. 1996;28(2):213–223.

[17] Fried R, Gather U. On rank tests for shift detection in time series. Computational Statistics & Data Analysis. 2007;52(1):221–233.

[18] Fried R. On the robust detection of edges in time series filtering. Computational Statistics & Data Analysis. 2007;52(2):1063–1074.

[19] Wilcox RR. Applying contemporary statistical techniques. Amsterdam: Academic Press; 2003.

[20] Maronna RA, Martin RD, Yohai VJ. Robust statistics: Theory and methods. Chichester: Wiley; 2006.

[21] Serfling R. Asymptotic relative efficiency in estimation. In: Lovric M, editor. International encyclopedia of statistical science. Berlin: Springer; 2011. p. 68–72.

[22] Hollander M, Wolfe DA. Nonparametric statistical methods. 2nd ed. New York: Wiley; 1999.

[23] Hájek J. A course in nonparametric statistics. San Francisco: Holden-Day; 1969.

[24] Ernst MD. Permutation methods: A basis for exact inference. Statistical Science. 2004;

19(4):676–685.

[25] Frisén M. Statistical Surveillance. Optimality and methods. International Statistical Review. 2003;71(2):403–434.

[26] Page ES. Continuous Inspection Schemes. Biometrika. 1954;41(1-2):100–115.

[27] Basseville M, Nikiforov IV. Detection of abrupt changes: Theory and application. Englewood Cliffs: Prentice Hall; 1993.

[28] Gan FF. An optimal design of EWMA control charts based on median run length. Journal of Statistical Computation and Simulation. 1993;45(3-4):169–184.

[29] Morgenthaler S, Tukey JW. Configural polysampling: A route to practical robustness. New York: Wiley; 1991.

[30] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria. 2014; Available from: `http://www.R-project.org/`.

[31] Bischl B, Lang M, Mersmann O. Batchexperiments: Statistical experiments on batch computing clusters.. 2014; R package version 1.4; Available from: `http://CRAN.R-project.org/package=BatchExperiments`.

[32] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria. 2015; Available from: `http://www.R-project.org/`.

[33] Wickham H. ggplot2: elegant graphics for data analysis. Springer New York; 2009.

[34] Sharpsteen C, Bracken C. tikzdevice: R graphics output in latex format. 2015; R package version 0.8.1; Available from: `http://CRAN.R-project.org/package=tikzDevice`.

[35] Høyland A. Robustness of the Hodges-Lehmann estimates for shift. The Annals of Mathematical Statistics. 1965;36(1):174–197.

[36] Schettlinger K, Fried R, Gather U. Real-time signal processing by adaptive repeated median filters. International Journal of Adaptive control and Signal Processing. 2010;24(5):346–362.

Appendix A. Figures

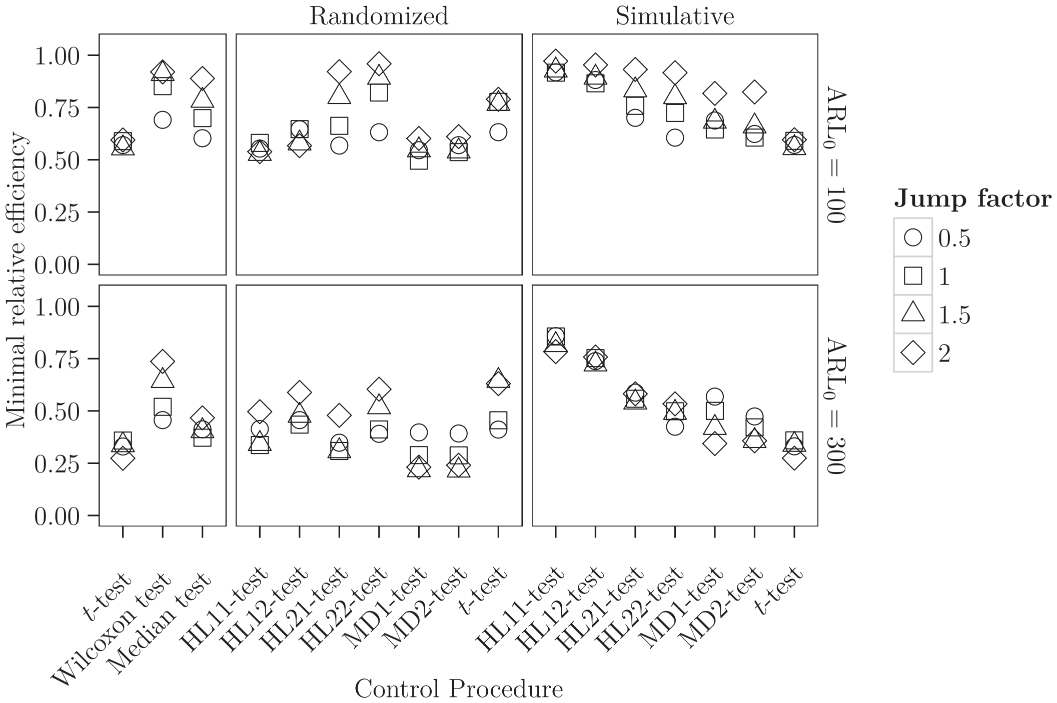


Figure A1.: Minimal relative efficiency based on the ARL for each control method over all considered distributions depending on the jump height.

Appendix B. Tables

Table B1.: Minimal MRL_1 achieved for the different error distributions split by jump height and fixed ARL_0 over all control procedures.

		Jump factor Δ			
ARL_0	Distribution	0.5	1	1.5	2
100	$\mathcal{N}(0,1)$	41	29	39	26
	t_5	42	29	28	26
	t_2	39	28	26	26
	χ^2_3	31	26	25	24
300	$\mathcal{N}(0,1)$	90	46	28	27
	t_5	90	41	28	27
	t_2	88	32	28	27
	χ^2_3	31	26	25	24