Bernoulli **24**(1), 2018, 526–564
DOI: 10.3150/16-BEJ887

A MOSUM procedure for the estimation of multiple random change points

BIRTE EICHINGER[1] and CLAUDIA KIRCH[2]

[1] *Karlsruhe Institute of Technology (KIT), Institute of Stochastics, Kaiserstr. 89, 76133 Karlsruhe, Germany. E-mail: birte.muhsal@kit.edu*
[2] *Otto-von-Guericke University Magdeburg, Institute for Mathematical Stochastics, Postfach 4120, 39106 Magdeburg, Germany. E-mail: claudia.kirch@ovgu.de*

In this work, we investigate statistical properties of change point estimators based on moving sum statistics. We extend results for testing in a classical situation with multiple deterministic change points by allowing for random exogenous change points that arise in Hidden Markov or regime switching models among others. To this end, we consider a multiple mean change model with possible time series errors and prove that the number and location of change points are estimated consistently by this procedure. Additionally, we derive rates of convergence for the estimation of the location of the change points and show that these rates are strict by deriving the limit distribution of properly scaled estimators. Because the small sample behavior depends crucially on how the asymptotic (long-run) variance of the error sequence is estimated, we propose to use moving sum type estimators for the (long-run) variance and derive their asymptotic properties. While they do not estimate the variance consistently at every point in time, they can still be used to consistently estimate the number and location of the changes. In fact, this inconsistency can even lead to more precise estimators for the change points. Finally, some simulations illustrate the behavior of the estimators in small samples showing that its performance is very good compared to existing methods.

Keywords: binary segmentation; change point; hidden Markov model; moving sum statistics; regime switching model

1. Introduction

There are essentially two approaches to multiple change point problems: Model selection and hypothesis testing. Model selection was first proposed by [44] who used Schwarz's criterion [40] for estimating the number of changes in the mean in an otherwise independent and normally distributed sequence of random variables. More recently, approaches based on least squares ([45] in context of mean changes or [31] for multivariate regression models), least absolute deviations ([7] for mean and simultaneous variance changes or [3] for linear regression models) or the minimum description length (confer [13] for linear autoregressive models) have been proposed. Since such procedures are usually computationally expensive, there exist many papers concerned with solving the optimization problem in an efficient manner for example, [23]. A different approach via an application of the Lasso was proposed by [20] resulting in an algorithm in linear time. Chen, Gupta and Pan [9] and Pan and Chen [36] refine existing information criteria such as the Schwarz criterion by making the model complexity not only a function of the number of change points but also a function of the location of the change points. This was motivated by the

1350-7265 © 2018 ISI/BS

unnecessary complexity of the model, when structural breaks appear close to each other, at the beginning or at the end of the data set.

Most papers concerned with hypothesis testing in the context of change point problems propose tests for a fixed or bounded number of changes (confer [4] for such F-type tests or [32] for a generalization to M-tests), where a large class of literature is concerned with the at most one change alternative (confer e.g., the monograph by [11]). Tests designed for at most one change still have some power for multiple changes, which is the idea behind binary segmentation methods first proposed by [43], where a change point is estimated if the test is significant and then the procedure is repeated for each segment until it is no longer significant. While binary segmentation procedures are adaptable to different change point problems and can be coded in a computational efficient way, it can be difficult to interpret the results in terms of significance due to the multiple testing involved. Furthermore, the power can suffer greatly under certain configurations of multiple changes, a drawback that is overcome by introducing an additional randomization step (to select the segment to be tested) proposed by [19]. Another recent fully parametric approach [18] minimizes the number of change points over the acceptance region of a suitable multiscale test. Multiscale tests have also been proposed by [15], which can be used to test the null hypothesis of stationarity versus possible change point alternatives.

In the context of testing, moving sum (MOSUM) or scan statistics have already been investigated by [5,10] or [21] as well as [22]. More recently, [39] use moving sum statistics in the frequency domain to detect changes in the autocovariance structure of multivariate time series. By construction, these procedures control the overall significance level thus avoiding issues due to multiple testing.

In this paper, we investigate properties of estimators for the change point location based on such statistics, an idea proposed but not mathematically analyzed by [2]. This method does not require to fix an upper bound for the number of changes and is computationally cheap, in fact the proposed method for i.i.d. errors achieves linear complexity.

The aim of this paper is twofold: First, the theoretic properties of these estimators are derived showing consistency for the number and location of the change points. Furthermore, rates of convergence for these estimators are obtained. The corresponding asymptotic distribution in a special case shows that these rates cannot in general be improved.

The second aim of this paper is to investigate the behavior of these estimators in the presence of random change points, as they occur for example, in regime switching models (for a recent survey we refer to [17]), which are used to model structural breaks in time series. In these models, a non-observable process $\{Q_i : i \in \mathbb{N}\}$ governs the regime of the time series and a change point occurs when this non-observable process $\{Q_i\}$ switches to another state. As long as this non-observable process is independent of the error sequence governing the stationary regimes, the corresponding samples of such a regime switching model look as if they were from a classical multiple change point model with deterministic (but unknown) change points. For special cases such as Hidden Markov Models, one can make use of the structure of the unobservable process to reconstruct the sequence of states by use of for example, the Viterbi algorithm (confer [8]). Nevertheless, it is of interest whether asymptotic results for the classical change point setting carry over to situations with random changes, as they offer a different nonparametric approach – not making explicit assumptions on the structure of the unobservable process (except that switches are rare) – to reconstruct the change points and in a second step the states.

The paper is organized as follows: In the next section, we introduce moving sum statistics which have already been considered in the literature in the context of testing for deterministic (but unknown) change points. In the first subsection, we introduce the multiple change model including possibly random change points that we will be using throughout the paper and compare it with the classical multiple change situation. In Section 2.2, we give the distribution of the moving sum statistic if no change points are present and show that the corresponding tests have asymptotic power one in this more general setting. Because the small sample performance of both tests and estimators depends crucially on the estimator for the (long-run) variance, we propose to use a new moving sum estimator in Section 2.3. In Section 3, we explain how the above moving sum statistics can be used to consistently estimate both the number of change points as well as their locations, derive rates in Section 3.2 for the estimators of the locations and show that those rates are strict in Section 3.3. In Section 3.4, some non-asymptotic results for i.i.d. innovations with known variance are given. In Section 3.5, we discuss the problems of bandwidth choice inherent to the procedure. Some simulations in Section 4 illustrate the small sample behavior of these estimators, before the proofs are given in Section 5.

2. MOSUM tests for multiple changes

In this section, the multiple mean change problem is introduced as well as the MOSUM statistic in the context of testing. Furthermore, we introduce a new moving sum (long-run) variance estimator which increases the power under alternatives and leads to more precise estimators in small samples.

2.1. Modeling multiple changes

The classical change point model, which allows for multiple changes in the mean, is defined by

$$X_i = \sum_{j=1}^{q+1} \mu_j 1_{\{k_{j-1,n} < i \le k_{j,n}\}} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where

$$0 = k_{0,n} < k_{1,n} = \lfloor \vartheta_1 n \rfloor \le \cdots \le k_{q,n} = \lfloor \vartheta_q n \rfloor \le k_{q+1,n} = n, \qquad 0 < \vartheta_1 \le \ldots \le \vartheta_q \le 1.$$

The number of structural breaks $q \in \mathbb{N}$, the change points $k_{1,n}, \ldots, k_{q,n}$ as well as the expected values $\mu_1, \ldots, \mu_{q+1} \in \mathbb{R}$ with $\mu_j \neq \mu_{j+1}$, $j = 1, \ldots, q$, are unknown and the centered stationary error sequence $\varepsilon_1, \ldots, \varepsilon_n$ fulfills conditions stated below.

In this paper, we investigate the theoretic properties of estimators based on moving sum statistics for the number of changes $q = q_n$ as well as the locations of the change points $k_{j,n}$. We will not only show that they are consistent but also derive the rates of convergence and in the above special case the joint asymptotic distribution which shows that these rates are strict in general. While in the above classical model, the distance between change points is proportional to n, we only require that it is larger than a multiple of the bandwidth (see Assumption A.2).

The sample paths of this classical change point model coincide with sample paths for many regime-switching models, where the change points (as well as the number of change points) are random rather than deterministic. Consequently, one can expect methods for the above classical change point problem to yield consistent results also for regime-switching models, which could help to analyze their properties. For this reason, we investigate the properties of our estimation procedure for the more general situation of random change points which includes the following regime-switching models:

$$X_i = \tilde{\mu}_{Q_i} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

with possible (conditional) expectations $\tilde{\mu}_1, \ldots, \tilde{\mu}_K \in \mathbb{R}$, where $\tilde{\mu}_i \neq \tilde{\mu}_j$ for $i, j = 1, \ldots, K$, $i \neq j$, and errors $\varepsilon_1, \ldots, \varepsilon_n$ fulfilling the conditions below. The conditional expectation of X_i is determined by a non-observable $\{1, \ldots, K\}$-valued stationary process $\{Q_i : 1 \leq i \leq n\}$, $K \in \mathbb{N}$, where we can even allow for an infinite state space. The key features of a regime switching model for which change point methods are applicable are (a) independence of the changes (i.e., of $\{Q_i\}$) and the error sequence and (b) long duration times of the non-observable process $\{Q_i\}$. Consequently, observation switching models (confer [29]) such as threshold models introduced by [42] or i.i.d. switching models are not covered by our theory in contrast to Hidden Markov models where $\{Q_i\}$ is independent of the error sequence and with a low tendency to switch. For a fixed realization the number of time points between two adjacent change points is fixed and finite, so that those changes become invisible asymptotically if naive asymptotics are used. Similarly to locally stationary processes [12], we adopt a different approach for asymptotics, assuming an underlying triangular scheme

$$X_i^{(n)} = \tilde{\mu}_{Q_i^{(n)}} + \varepsilon_i^{(n)}$$

with a sequence of non-observable processes $\{Q_i^{(n)} : 1 \leq i \leq n\}$. Furthermore, we consider both a setup with a fixed (or at least bounded) number of change points $q_n = q$ ($q_n = O(1)$) as well as a setup with an increasing number of change points $q_n \to \infty$, where the distance between adjacent change points grows at least with the same rate as the bandwidth of the moving sum statistic (see Assumption A.2 below). In small samples, this translates to a low tendency to switch and only relatively few change points compared to the sample size.

We will use the following reparametrization of the above regime switching model which emphasizes the similarities to the classical multiple change situation (where again $k_{0,n} = 0$ and $k_{q_n+1,n} = n$)

$$X_i = \sum_{j=1}^{q_n+1} \mu_{j,n} 1_{\{k_{j-1,n} < i \leq k_{j,n}\}} + \varepsilon_i, \qquad \mu_{j,n} \in \{\tilde{\mu}_{1,n}, \ldots, \tilde{\mu}_{K_n,n}\},$$

$i = 1, \ldots, n$, where in contrast to the classical multiple change model the number of change points q_n, the change points $k_{1,n}, \ldots, k_{q_n,n}$ as well as $\mu_{1,n}, \ldots, \mu_{q_n+1,n}$ are random variables. Due to the triangular scheme the classical multiple change model is a special case in this setting.

In the following, it will also be useful to rewrite the model in terms of mean differences as

$$X_i = \mu_{1,n} + \sum_{j=1}^{q_n} d_{j,n} 1_{\{i>k_{j,n}\}} + \varepsilon_i, \qquad d_{j,n} = \mu_{j+1,n} - \mu_{j,n}, \tag{2.1}$$

where $d_{j,n}$ are the mean changes. In the following, we will suppress the dependence of $k_{j,n}$, $d_{j,n}$ and $\mu_{j,n}$ on n and simply write k_j, d_j and μ_j for notational ease.

We are now ready to state the assumptions on the error distribution.

Assumption A.1.

(a) *The number and magnitude of changes and their locations* $\{q_n, d_{j,n}, k_{j,n}, j = 1, \dots, q_n\}$ *are independent of the error sequence* $\{\varepsilon_i\}$.

(b) *There exists a standard Wiener process* $\{W(k) : 1 \le k \le n\}$ *and* $\nu > 0$ *such that* (*possibly after changing the probability space*)

$$\sum_{i=1}^{n} \varepsilon_i - \tau W(n) = O\big(n^{1/(2+\nu)}\big) \qquad a.s.$$

with an existing and strictly positive long-run variance

$$\tau^2 = \sigma^2 + 2\sum_{h>0} \gamma(h) > 0, \qquad \gamma(h) = \operatorname{cov}(\varepsilon_0, \varepsilon_h), \qquad \sigma^2 = \operatorname{var}(\varepsilon_1).$$

(c) *For some* $\gamma > 2$ *and some constant* $C > 0$ *it holds for any* $-\infty < \ell \le u < \infty$

$$\mathrm{E}\left|\sum_{i=\ell}^{u} \varepsilon_i\right|^{\gamma} \le C|u - \ell + 1|^{\gamma/2}.$$

Part (a) is needed to guarantee that the stochastic behavior of (partial) sums of the error sequence does not depend on the change points. This guarantees for example that the moving sum statistic behaves like under the null hypothesis of no change if the next change point is sufficiently far away. Since all the quantities in (a) are deterministic in the classical change point setting, Assumption A.1 is automatically fulfilled in this case.

Invariance principles as in (b) have first been derived for i.i.d. random variables by [26] and [27]. Subsequently, many classes of time series have also been considered, for example, mixing time series (Theorem 4 in [28]) or near-epoch dependent time series [30].

The assumption in (c) is needed to get rates of convergence for the change point estimators. It holds for example, for independent random variables or linear processes, for which it follows from the Beveridge–Nelson decomposition (confer [6,37]). It also holds for martingale difference sequences, certain Φ-mixing sequences (confer [41], Theorem 3.7.8) but also for certain α-mixing sequences (confer [47], Theorem 1).

2.2. Testing for multiple mean changes using a moving sum statistic

Hušková and Slabý [22] proposed to use moving sum (MOSUM) statistics for testing the classical multiple change hypothesis in i.i.d. data. While in this paper, the focus lies on the properties of corresponding estimators for the number and location of changes, the asymptotic null distribution plays an important role in this analysis. Therefore, we obtain the null asymptotics in this section for dependent errors fulfilling Assumption A.1(b). Furthermore, we show that the corresponding test statistic has asymptotic power one not only in the classical setting but also for the more general random change model above. Consider the following moving sum statistic

$$\begin{aligned} T_n(G) &= \max_{G\le k\le n-G} \frac{|T_{k,n}(G)|}{\tau}, \\ T_{k,n}(G) &= T_{k,n}(G; X_1,\dots,X_n) = \frac{1}{\sqrt{2G}}\left(\sum_{i=k+1}^{k+G} X_i - \sum_{i=k-G+1}^{k} X_i\right), \end{aligned} \tag{2.2}$$

with bandwidth $G = G(n)$ fulfilling

$$\frac{n}{G} \longrightarrow \infty \quad \text{and} \quad \frac{n^{\frac{2}{2+\nu}}\log n}{G} \longrightarrow 0, \tag{2.3}$$

where ν and τ are as in Assumption A.1(b). The bandwidth assumption guarantees that G converges to infinity but not too fast. We can calculate the above statistic in linear time by using a simple recursive calculation

$$\sqrt{2G}T_{k+1,n}(G) = \sqrt{2G}\big(T_{k,n}(G) + X_{k-G+1} - 2X_{k+1} + X_{k+G+1}\big). \tag{2.4}$$

The statistic in (2.2) compares at every time point $G \le k \le n-G$ the mean of the subsample $X_{k-G+1},\dots,X_k$ with the mean of the subsample $X_{k+1},\dots,X_{k+G}$, where a large difference indicates a change at this point. At point k, this is essentially the Wald version of the likelihood ratio statistic for the sample $X_{k-G+1},\dots,X_{k+G}$ with a possible change at k [25].

Formally, we test the null hypothesis of no change $H_0 : q_n = 0$, or equivalently $H_0 : X_i = \mu_1 + \varepsilon_i$, $i = 1,\dots,n$, versus the alternative that at least one change occurs $H_1 : q_n \ge 1$.

The following theorem gives the null asymptotics of the test statistic quantifying the acceptable deviation from zero.

Theorem 2.1. *Let the null hypothesis hold, that is,* $X_i = \mu_1 + \varepsilon_i$, $i = 1,\dots,n$, *with* $\{\varepsilon_i\}$ *fulfilling Assumption* A.1(b). *If the bandwidth G fulfills* (2.3), *then*

$$a(n/G)T_n(G) - b(n/G) \xrightarrow{\mathcal{D}} \Gamma,$$

where Γ *follows a Gumbel extreme value distribution, that is,* $P(\Gamma \le x) = \exp(-2\exp(-x))$ *and*

$$a(x) = \sqrt{2\log x}, \qquad b(x) = 2\log(x) + \frac{1}{2}\log\log x + \log(3/2) - \frac{1}{2}\log\pi.$$

The assertion remains true if τ in $T_n(G)$ is replaced by an estimator $\hat{\tau}_{k,n}$, which can depend on the position k, and fulfills under the null hypothesis

$$\max_{G\le k\le n-G}\left|\hat{\tau}_{k,n}^2-\tau^2\right|=o_P\big(\big(\log(n/G)\big)^{-1}\big). \tag{2.5}$$

Consequently, we get an asymptotic level α test if the null hypothesis is rejected for $T_n(G)>D_n(G;\alpha)$ with

$$D_n(G;\alpha)=\frac{b(n/G)+c_\alpha}{a(n/G)},\qquad c_\alpha=-\log\log\frac{1}{\sqrt{1-\alpha}}. \tag{2.6}$$

We now turn to the asymptotic behavior of the test statistic under the alternative of at least one change, where we allow for random changes as in (2.1). The next theorem proves that this test rejects with asymptotic power one, while subsequent paragraphs derive the asymptotic properties of corresponding estimators for the location of the change points. To this end, we require the following assumptions on the distance between change points as well as the magnitude of changes.

Assumption A.2. *It holds for the distance between two changes*

$$P\Big(\min_{0\le j\le q_n}|k_{j+1}-k_j|>cG\Big)\to 1$$

for some $c\ge 2$ specified below.

This assumption shows the connection between the distance between changes and the choice of bandwidth G. It states that the distance between two change points grows at least like G where empirical evidence suggests that G should be chosen as large as possible but such that any window of length $2G$ contains at most one change. On the other hand, the distance between change points can be of smaller order than n, which is different from the classical change point problem but of importance for the analysis in situations, where changes tend to be closer together (as for the regime-switching model).

Assumption A.3. *It holds for the magnitude of changes*:

$$\text{(a)}\quad \frac{1}{\min_{1\le j\le q_n}d_j^2}=o_P\left(\frac{G}{\log(n/G)}\right).$$

$$\text{(b)}\quad P\Big(\min_{1\le j\le q_n}|d_j|\ge\delta_n\Big)\to 1$$

for some sequence $\delta_n>0$ indicating a non-stochastic lower bound for the occurring mean changes.

Both assumptions give lower bounds for the mean changes and allow in particular for local changes (if $\delta_n\to 0$). They will play an important role in the derivation of rates of convergence for the proposed change point estimators.

The estimator for the long-run variance does not need to be uniformly consistent under alternatives (as in (2.5)) but the following restrictions apply.

Assumption A.4.

(a) *The* (*long-run*) *variance estimator is strictly positive with probability tending to one*,

$$P\Big(\min_{G\le k\le n-G}\hat{\tau}_{k,n}>0\Big)\to 1.$$

(b) *The* (*long-run*) *variance estimator fulfills*

$$\max_{G\le k\le n-G}\hat{\tau}^2_{k,n}=o_P\left(\frac{G\min_{1\le j\le q_n}d_j^2}{\log(n/G)}\right).$$

(c) *The* (*long-run*) *variance estimator is consistent outside a* $2G$*-neighborhood of change points with rate* $\log(n/G)^{-1}$, *i.e.*

$$\max_{|k-k_j|\ge G,j=1,\dots,q_n}\big|\hat{\tau}^2_{k,n}-\tau^2\big|=o_P\big(\big(\log(n/G)\big)^{-1}\big).$$

In the next section, we propose estimators for the long-run variance which fulfill the above assumptions for an appropriate bandwidth choice.

Some long-run variance estimators such as the below flat-top kernel estimator do not guarantee that the corresponding estimate is always positive. Under the null hypothesis this effect vanishes asymptotically due to consistency of the estimator. However, under alternatives this is not guaranteed which is why we need to make Assumption A.4(a) to ensure that the estimator is neither negative nor zero asymptotically. It does not matter for the procedure if τ is very small or even converges to zero (while being strictly positive for all n) as this only results in a larger value of the statistic, which increases the chance of detection keeping in mind that Assumption A.4(c) guarantees consistency of the behavior away from change points. Typically, estimators for the long-run variance that allow for negative values are modified in a way that ensures Assumption A.4(a) as negative variance estimates cause problems in almost all statistical procedures. The standard solution for the flat-top kernel estimator is to truncate the estimator from below by the sample variance divided by the logarithm of the sample size to guarantee positivity, scale invariance in addition to asymptotic consistency if no changes are present. In this sense Assumption A.4(a) is merely a technicality.

Assumption A.4(b) is a restriction on the rate of divergence for the (long-run) variance estimator under alternatives. It follows from Assumptions A.3(a) and A.4(c) under the null hypothesis respectively away from changes.

Assumption A.4(c) holds for example, for translation invariant estimators, that is, $\hat{\tau}^2_{k,n}=\hat{\tau}^2_{k,n}(X_{k-G+1},\dots,X_{k+G})=\hat{\tau}^2_{k,n}(X_{k-G+1}-\mu,\dots,X_{k+G}-\mu)$ for any μ, if additionally

$$\max_{G\le k\le n-G}\big|\hat{\tau}^2_{k,n}(\varepsilon_{k-G+1},\dots,\varepsilon_{k+G})-\tau^2\big|=o_P\big(\big(\log(n/G)\big)^{-1}\big).$$

Theorem 2.2. *Let* $X_1,\dots,X_n$ *follow* (2.1) *and let Assumption* A.2 *with* $c=2$ *and Assumption* A.3(a) *hold. Furthermore, let* $\{\varepsilon_t\}$ *fulfill Assumption* A.1(a) *and* (b). *Then, we get for any* $z\in\mathbb{R}$

$$P\big(a(n/G)T_n(G)-b(n/G)\ge z\big)\to 1,$$

that is, the test has asymptotic power one. The assertion remains true, if τ is replaced by an estimator $\hat{\tau}_{k,n}$ which can depend on k such that Assumption A.4(a) *and* (b) *hold.*

2.3. Estimating the asymptotic variance

The small sample power as well as precision of estimators depends crucially on how the (long-run) variance τ^2 is estimated. While using a standard estimator based on the full sample yields best results under the null hypothesis, this estimator is contaminated by the mean changes under alternatives such that the estimate is too large (confer Figure 1). While the corresponding test has still asymptotic power one under relatively mild assumptions (confer Theorem 2.2), it will suffer a great power loss in small samples compared to a test, where the true (long-run) variance is used. In the at most one change situation, the standard solution for classic CUSUM (cumulative sum) statistics is to estimate the possible change point $\hat{k}_n$ (using the point of maximum of the statistic) and then to calculate the sample variance after a separate mean correction for both subsamples

$$\frac{1}{n}\sum_{i=1}^{\hat{k}_n}\left(X_i - \frac{1}{\hat{k}_n}\sum_{j=1}^{\hat{k}_n} X_j\right)^2 + \frac{1}{n}\sum_{i=\hat{k}_n+1}^{n}\left(X_i - \frac{1}{n-\hat{k}_n}\sum_{j=\hat{k}_n+1}^{n} X_j\right)^2,$$

where $\hat{k}_n$ is a suitable estimator for the possible change point. Since the number of changes is unknown in our case, such an approach is no longer feasible. To avoid this problem, we propose to use a time dependent MOSUM type estimator which treats each time point k as a possible change point estimating the variance after a separate mean correction for the segment before and

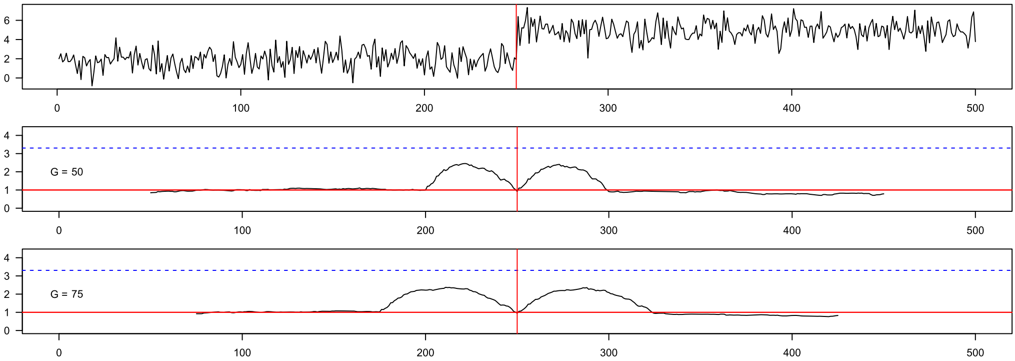

Figure 1. Performance of the variance estimators $\hat{\sigma}_{k,n}^2$ (black solid line) as in (2.7) and the empirical variance $\hat{\sigma}_n^2$ as in (2.8) (without taking possible change points into account) (dashed line). The true variance of 1 is indicated by the horizontal red line.

after k. Precisely, we propose to use the following variance estimator in case of i.i.d. errors:

$$\hat{\sigma}_{k,n}^2 := \frac{1}{2G}\left(\sum_{i=k-G+1}^{k}(X_i - \overline{X}_{k-G+1,k})^2 + \sum_{i=k+1}^{k+G}(X_i - \overline{X}_{k+1,k+G})^2\right),$$

$$\text{where } \overline{X}_{l,j} = \frac{1}{j-l+1}\sum_{i=l}^{j} X_i, \tag{2.7}$$

which can be calculated in linear time similarly to (2.4). Figure 1 illustrates the behavior of this estimator in comparison to the sample variance

$$\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \tag{2.8}$$

for the time series in the upper picture which includes one change point (indicated by the vertical line). The two lower pictures show the sample variance (dashed horizontal line) as well as the time dependent performance of estimator (2.7) (solid line) for different choices of bandwidths, where the true variance is 1 (as indicated by the red horizontal line). While the sample variance clearly overestimates the variance, the time dependent MOSUM estimator is consistent in regions with no change as well as right at the change point but it overestimates the variance close to a change but not right at it. While this later feature may first seem unattractive, we will explain in Section 4 why this is a desirable property that can even help to get more precise estimates for the location of the change points in small samples.

Similarly, we propose to use time dependent MOSUM versions of flat-top kernels as proposed by [38] for the long-run variance in case of dependent errors

$$\hat{\tau}_{k,n}^2 := \hat{\gamma}_k(0) + 2\sum_{h=1}^{\Lambda_n}\omega(h/\Lambda_n)\hat{\gamma}_k(h)$$

with autocovariance estimator

$$\begin{aligned}\hat{\gamma}_k(h) := &\frac{1}{2G}\sum_{i=k-G+1}^{k-h}(X_i - \overline{X}_{k-G+1,k})(X_{i+h} - \overline{X}_{k-G+1,k})\\ &+ \frac{1}{2G}\sum_{i=k+1}^{k+G-h}(X_i - \overline{X}_{k+1,k+G})(X_{i+h} - \overline{X}_{k+1,k+G}),\end{aligned}$$

bandwidth Λ_n and suitable weights ω. For example, Bartlett weights defined by $w(x) = (1-x)_+$ or the following flat-top weights can be used

$$w(t) = \begin{cases} 1, & |t| \le 1/2, \\ 2\big(1-|t|\big), & 1/2 < |t| < 1, \\ 0, & |t| \ge 1. \end{cases}$$

Theorem 2.3. *Let $X_i = \mu + \varepsilon_i$, $i = 1, \dots, n$, $\{\varepsilon_i\}$ fulfill Assumption* A.1(b) *as well as* $\mathrm{E}\,|\varepsilon_i|^4 < \infty$ *and*

$$\sup_{h\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} \sum_{l\in\mathbb{Z}} \big|\nu(h,k,l)\big| < \infty, \tag{2.9}$$

where $\nu(h,r,s) = \operatorname{cov}(\varepsilon_1\varepsilon_{1+h}, \varepsilon_{1+r}\varepsilon_{1+s})$. *Then it holds*:

(a) *If* $n/G^2 = O(1)$, *we get*

$$\max_{G\le k\le n-G} \big|\hat{\sigma}^2_{k,n} - \sigma^2\big| = O_P\left(\frac{n^{1/2}}{G}\right).$$

(b) *If* $\Lambda_n^2 n/G^2 = O(1)$ *and the weights fulfill* $0 \le w(x) \le C$, *then*

$$\max_{G\le k\le n-G} \big|\hat{\tau}^2_{k,n} - \tau^2\big| = O_P\left(\frac{\Lambda_n n^{1/2}}{G} + r_n\right),$$

where

$$r_n = \sum_{h\in\mathbb{Z}} \big|w(h/\Lambda_n) - 1\big|\big|\gamma(h)\big|.$$

Remark 2.1. (a) In (a), we get the weaker rate $O_P(n^{2/(2+\nu)}/G)$ if only $\mathrm{E}\,|\varepsilon_i|^{2+\nu} < \infty$, $0 < \nu < 2$, and the stronger rate $O_P(\sqrt{\log(n/G)/G})$ for $\nu > 2$ (for details we refer to [35], proof of Theorem 6.14).

(b) The rate r_n depends on the choice of weights $w(\cdot)$ in addition to the convergence rate of $\gamma(h)$. If $\sum_{h\in\mathbb{Z}} h^\alpha |\gamma(h)| < \infty$, $0 < \alpha \le 1$, which is a standard assumption in time series analysis, then for both the Bartlett as well as flat-top weights $r_n = O(\Lambda_n^{-\alpha})$.

From Theorem 2.3 we get (2.5) for i.i.d. innovations if $n^{1/2}\log(n/G)/G \to 0$ which is implied by (2.3) if $\nu \le 2$. For dependent errors, we obtain (2.5) under additional assumptions on the bandwidth possibly in addition to the autocovariance structure as indicated by Remark 2.1(b).

If changes are present, the estimators are influenced by these changes as demonstrated in Figure 1, but we still obtain the following theorem.

Theorem 2.4. *Let the error sequence $\{\varepsilon_i\}$ fulfill Assumption* A.1(b), $\mathrm{E}\,|\varepsilon_1|^4 < \infty$, (2.9). *Furthermore let Assumption* A.2 *hold with $c = 2$ in addition to* $\max_{1\le j\le q_n} |d_{j,n}| = O_P(1)$, *i.e. the mean changes are bounded. Then, we get for the random change point model* (2.1) (*which includes the classical model*):

(a) *If* $n/G^2 \to 0$,

$$\max_{G\le k\le n-G} \hat{\sigma}^2_{k,n} = O_P(1).$$

(b) *If* $\Lambda_n^2 n/G^2 \to 0$,

$$\max_{G\le k\le n-G} \hat{\tau}_{k,n}^2 = O_P(\Lambda_n).$$

3. MOSUM-based estimators for multiple change points

Additionally to testing for changes, the MOSUM statistic can be used to estimate the number and location of the change points (in rescaled time). In this section, we will introduce those estimators and derive their asymptotic consistency. Additionally, we will obtain rates of convergence for the estimator of the locations of the changes and show that these rates are strict in general.

3.1. Asymptotic consistency

Figure 2 shows a time series with i.i.d. errors and three change points marked by vertical lines (upper panel) as well as the statistic $T_{k,n}(G)/\hat{\tau}_{k,n}$ as a function of k (lower panel), where the horizontal line marks the asymptotic critical value at the 5% level and $\hat{\tau}_{k,n} = \hat{\sigma}_{k,n}$ as in (2.7). Obviously, the test statistic exceeds the critical value in intervals around the true change points with the local maxima close to the locations of the change points. The calculations of these estimators can also be done in linear time (given the MOSUM process).

Based on this observation, we define estimators for the number of change points as well as their locations as follows: Consider all pairs of indices (v_j, w_j) such that

$$\begin{aligned} &\hat{\tau}_{k,n}^{-1}\left|T_{k,n}(G)\right| \ge D_n(G;\alpha_n) \qquad \text{for } k = v_j,\ldots,w_j,\\ &\hat{\tau}_{k,n}^{-1}\left|T_{k,n}(G)\right| < D_n(G;\alpha_n) \qquad \text{for } k = v_j-1, w_j+1, \\ &\qquad w_j - v_j \ge \eta G \qquad \text{with } 0<\eta<1/2 \text{ arbitrary but fixed,}\end{aligned} \tag{3.1}$$

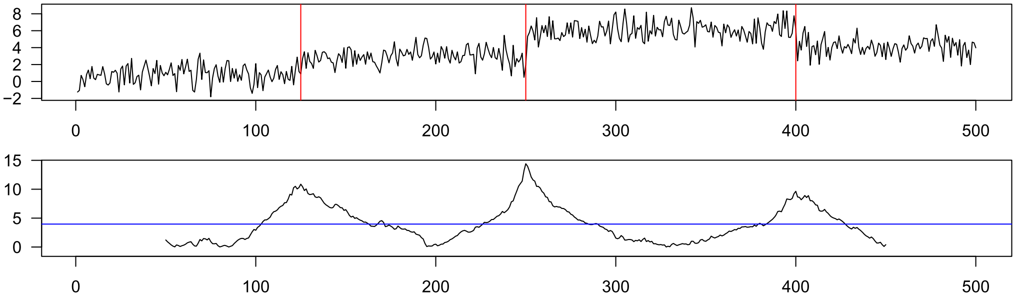

Figure 2. Time series with i.i.d. errors and three change points marked by the vertical lines (upper panel) as well as $\hat{\sigma}_{k,n}^{-1}|T_{k,n}(G)|$ (lower panel), where the horizontal line gives the asymptotic critical value at the 5% level.

where $D_n(G;\alpha)$ is as in (2.6). Then,

$$\hat{q}_n \widehat{=} \text{ the number of pairs } (v_j, w_j) \tag{3.2}$$

is an estimator of the number of change points q, while

$$\hat{k}_j := \arg \max_{v_j \le k \le w_j} \frac{|T_{k,n}(G)|}{\hat{\tau}_{k,n}} \tag{3.3}$$

are estimators for the locations of the change points.

Condition (3.1) is necessary to avoid overestimation by spurious local maxima exceeding the critical value on the boundary between significant and insignificant areas, that is, when the statistic crosses the critical line.

For these estimators to be consistent we require that the sequence α_n converges to 0 but not too fast, more precisely

$$\alpha_n \to 0, \qquad \frac{c_{\alpha_n}}{a(n/G)} = O(1), \tag{3.4}$$

where $a(\cdot)$ is as in Theorem 2.1 and c_{α_n} as in (2.6).

This method has already been proposed but not mathematically analyzed by [2]. A related but somewhat different approach based on local maxima of a MOSUM likelihood ratio statistic in addition to a model selection procedure for AR-time series has very recently been considered by [46]. Furthermore, [33] use a related MOSUM procedure for estimating change points in point processes.

We are now ready to prove that the number of change points as well as the locations in rescaled time are estimated consistently. In the next sections, we derive convergence rates of the change point estimators and show that these rates are strict in general.

Theorem 3.1. *Let $X_1, \dots, X_n$ follow* (2.1) *and let Assumption* A.2 *with $c = 2$ and Assumptions* A.3(a) *and* A.4 *hold. Furthermore, let $\{\varepsilon_t\}$ fulfill Assumption* A.1(a) *and* (b).

If α_n fulfills (3.4), *then it holds as $n \to \infty$*

$$\text{(a)} \quad P(\hat{q}_n = q_n) \longrightarrow 1.$$

$$\text{(b)} \quad P\Big(\max_{1 \le j \le q_n} |\hat{k}_j 1_{\{j \le \hat{q}_n\}} - k_j| \ge G\Big) \longrightarrow 0.$$

The above assertions are derived by using bounds for the type-I as well as type-II errors of the corresponding change point tests (see Lemma 5.1).

3.2. Convergence rates

By Assumption A.1(c) we get the following forward as well as backward Hájék–Rényi-type inequalities.

Lemma 3.1. *Under Assumption* A.1(c) *it holds for any positive and non-increasing sequence* $\{c_k\}$, *any* $1 \leq \ell \leq u$, *any* $m \in \mathbb{Z}$ *and any* $\delta > 0$

$$\text{(a)} \quad \delta^{\gamma} P\left(\max_{\ell \leq k \leq u} c_k \left| \sum_{j=m+1}^{m+k} \varepsilon_j \right| > \delta \right) \leq \widetilde{C}\left(c_{\ell}^{\gamma} \ell^{\gamma/2} + \sum_{k=\ell+1}^{u} c_k^{\gamma} k^{\gamma/2-1} \right),$$

$$\text{(b)} \quad \delta^{\gamma} P\left(\max_{\ell \leq k \leq u} c_k \left| \sum_{j=m-k+1}^{m} \varepsilon_j \right| > \delta \right) \leq \widetilde{C}\left(c_{\ell}^{\gamma} \ell^{\gamma/2} + \sum_{k=\ell+1}^{u} c_k^{\gamma} k^{\gamma/2-1} \right),$$

where $\widetilde{C}$ *only depends on* C *and* γ *of Assumption* A.1(c).

The following theorem gives detailed bounds for the deviation of the estimators of the change point locations from the true locations. Remark 3.1 simplifies these bounds in a way that yields standard assertions well known for the one-change-point situation and CUSUM statistics.

Theorem 3.2. *Let* $X_1, \ldots, X_n$ *follow* (2.1) *and let Assumption* A.2 *with* $c = 2$ *and Assumption* A.3 *hold and* $\hat{\tau}_{k,n}^2 = \hat{\tau}_n^2$ *fulfill Assumption* A.4. *Furthermore, let* $\{\varepsilon_t\}$ *fulfill Assumption* A.1 *and* α_n (3.4).

(a) *It holds for any* $1 \leq j \leq q_n$ *as well as* $1 \leq \xi_n \leq G$

$$P\left(|\hat{k}_j 1_{\{j \leq \hat{q}_n\}} - k_j| > \xi_n\right) = \delta_n^{-\gamma} \xi_n^{-\frac{\gamma}{2}} O(1) + o(1),$$

where the rates do not depend on j *and* δ_n *is as in Assumption* A.3(b).

(b) *If additionally* $P(q_n > \gamma_n) \to 0$, *then*

$$P\left(\max_{1 \leq j \leq q_n} |\hat{k}_j 1_{\{j \leq \hat{q}_n\}} - k_j| > \xi_n \right) = \gamma_n \delta_n^{-\gamma} \xi_n^{-\frac{\gamma}{2}} O(1) + o(1).$$

In both cases the additive term $o(1)$ stems from the fact that the assertions only hold on a sequence of asymptotic 1-sets (not depending on j). The bounds for these 1-sets can be made precise if the type-I and (in some sense) uniform type-II errors are known, similarly the $O(1)$ terms can be made precise using Lemma 3.1.

Remark 3.1.

(a) If the probability on the left-hand side of (a) is replaced by the conditional probability (given d_j), then δ_n on the right-hand side can be replaced by d_j.

(b) If $\delta_n^{-2} G^{-1} \to 0$, then

$$|\hat{k}_j 1_{\{j < \hat{q}_n\}} - k_j| = O_P\left(\delta_n^{-2}\right),$$

where the rate cannot be improved if d_j is asymptotically of the same rate as δ_n (confer also (a) above).

(c) If additionally the number of changes is stochastically bounded as in the classical change point situation, we obtain

$$\max_{1\le j\le q_n} |\hat{k}_j 1_{\{j<\hat{q}_n\}} - k_j| = O_P\big(\delta_n^{-2}\big),$$

where again the rate cannot in general be improved.

(d) Concerning the joint rate in a model with an increasing number of change points, we cannot derive the limit distribution of the maximum as in the next section, however, those rates cannot be improved in general. To this end, consider i.i.d. errors, $d_j = d$, $\hat{\tau}_{k,n} = \tau$ and q_n deterministic change points with $k_{j+1} - k_j > 4G$. On the set $M_n = \{\hat{q}_n = q_n, \max_{1\le j\le q_n} |\hat{k}_j - k_j| < G\}$ the estimators $\hat{k}_j$ only depend on the behavior of $X_{k_j-2G+1}, \dots, X_{k_j+2G-1}$, hence $(\hat{k}_j 1_{\{j<\hat{q}_n\}} - k_j)$, $j = 1, \dots, q_n$, are independent and identically distributed (on that set). Because M_n coincides with the asymptotic 1-set in (5.7) in the above setting (which does not depend on ξ_n), we get

$$\begin{aligned}
&P\Big(\max_{1\le j\le q_n} |\hat{k}_j 1_{\{j<\hat{q}_n\}} - k_j| > \xi_n\Big)\\
&\quad = 1 - \big(1 - P\big(|\hat{k}_1 - k_1| > \xi_n\big) + O\big(P\big(M_n^C\big)\big)\big)^{q_n} + O\big(P\big(M_n^C\big)\big)\\
&\quad = 1 - \exp\big[q_n \log\big(1 - P\big(|\hat{k}_1 - k_1| > \xi_n\big) + o(1)\big)\big] + o(1)\\
&\quad = 1 - \exp\big[-q_n(P\big(|\hat{k}_1 - k_1| > \xi_n\big) + o\big(P\big(|\hat{k}_1 - k_1| > \xi_n\big)\big) + o(1)\big] + o(1),
\end{aligned}$$

where the Taylor-expansion $\log(1-x) = x + o(1)$ (as $x \to 0$) has been used. Because the rate of $P(|\hat{k}_1 - k_1| > \xi_n)$ cannot be improved in general, neither can the joint rate even with an increasing number of change points.

Even though only local information is used in the estimation of the change points and the distance between change points is allowed to increase with a rate smaller than n (Assumption A.2), the rate of convergence for the change point estimator given in (b) above coincides with the optimal rates obtained in the at-most-one-change situation using the full sample (confer Theorem 2.8.2 in [11]). It also coincides with the optimal rate obtained in [45] (Theorem 1) in the classical mean change model with a fixed and known number of changes and independent errors and is better than the rate of convergence obtained in [20]. Furthermore, the rates are better than the ones obtained by [19].

Remark 3.2. In general the assertions of Theorem 3.2 do not carry over to situations, where an estimator $\hat{\tau}_{k,n}^2$ depending on k is used. For example, Assumptions A.4 do not rule out the possibility of the estimator $\hat{\tau}_{k,n}^2$ underestimating the variance (even asymptotically) close to the change point (but not at the change point), which would result in a distortion of the point of maximum away from the true change point. Consequently, a much more detailed analysis needs to be conducted in this case in order to obtain analogous results. This can be difficult in particular for estimators in the presence of dependence. A solution is to use the (long-run) variance estimator

$\hat{\tau}^2_{k,n}$ only to obtain $\hat{q}_n$ as well as the region (v_j, w_j) as in (3.1), but then replace (3.3) by

$$\hat{k}_j = \arg \max_{v_j \le k \le w_j} \left|T_{k,n}(G)\right|.$$

In this case, the proof of Theorem 3.2 remains correct and the assertions carry over.

On the other hand, the estimator $\hat{\sigma}^2_{k,n}$ as in (2.7) has the nice property of overestimating the variance close to (but not at) the change point, which should heuristically lead to an improvement for small samples and at least not to worse results asymptotically. The following corollary shows that this is in fact true for this particular estimator.

Corollary 3.1. *Let the assumptions of Theorem 3.2 be fulfilled for $\{\varepsilon_t\}$ i.i.d. with $\mathrm{E}\,|\varepsilon_1|^4 < \infty$ and use the variance estimator $\hat{\sigma}^2_{k,n}$ as in (2.7). If $\frac{n}{G^2} \to 0$, $\delta_n^{-2} G^{-1} \to 0$ and $d_j^2 = O_P(1)$, then it holds*

$$|\hat{k}_j 1_{\{j < \hat{q}_n\}} - k_j| = O_P\left(\delta_n^{-2}\right).$$

3.3. Asymptotic distribution

In this section, we derive the asymptotic distribution of the estimators for the locations for local changes. In particular, this shows that the convergence rates of the estimators are strict in general (confer also Remark 3.1). For non-local changes, the limit can be derived using analogous methods but it is no longer pivotal but depends on the underlying error distribution (for details in the classical at most one change situation for the CUSUM statistic, we refer to [1] as well as [14]).

In order to derive the asymptotic distribution, we need to make some stronger assumptions on the error sequence.

Assumption A.5. *Let the error sequence fulfill either* (i) *or* (ii):

(i) *$\{\varepsilon_i\}$ are i.i.d.*

(ii) *$\{\varepsilon_i\}$ are stationary and strong mixing and fulfill the functional central limit theorem in forward and backward time.*

Both assumptions ensure that functionals of the error sequence that are a fraction of G apart are asymptotically independent, where the mixing assumption can be relaxed if the asymptotic independence remains true. In particular, under Assumption A.2 with $c > 2$, the change point estimators are asymptotically independent. The forward functional central limit theorem is fulfilled by Assumption A.1(b), which follows for example for exponentially mixing sequences under moment conditions (confer [28]). Since mixing is a symmetric property, a backward invariance principle implying a backward functional central limit theorem holds under the same assumptions.

Theorem 3.3. *Let $X_1, \ldots, X_n$ follow (2.1) and let Assumption A.2 with $c > 2$ and Assumption A.3 hold and let $\hat{\tau}^2_{k,n} = \hat{\tau}^2_n$ fulfilling Assumption A.4. Furthermore, let $\{\varepsilon_t\}$ fulfill Assumption A.1 and α_n fulfills (3.4).*

(a) *If* $d_j \xrightarrow{P} 0$ *but* $d_j^2 G \xrightarrow{P} \infty$, *then it holds for* $j = 1, \dots, q_n$ *as* $n \to \infty$

$$\tau^{-2} d_j^2 (\hat{k}_j 1_{\{j \le \hat{q}_n\}} - k_j) \xrightarrow{\mathcal{D}} \arg\max\{W_s - |s|/\sqrt{6} : s \in \mathbb{R}\},$$

where $\{W_s : s \in \mathbb{R}\}$ *is a standard Wiener process.*

(b) *If additionally Assumption* A.5(i) *or* (ii) *holds, then for a fixed number of changes* $q_n = q$ *and* $\max_{j=1,\dots,q} d_j^2 \xrightarrow{P} 0$, *it holds as* $n \to \infty$

$$\tau^{-2}\big(d_1^2(\hat{k}_1 1_{\{1 \le \hat{q}_n\}} - k_1), \dots, d_q^2(\hat{k}_q 1_{\{q \le \hat{q}_n\}} - k_q)\big) \xrightarrow{\mathcal{D}} (S_1, \dots, S_q),$$

where

$$S_i = \arg\max\{W_s^{(i)} - |s|/\sqrt{6} : s \in \mathbb{R}\}$$

and $\{W_s^{(i)} : s \in \mathbb{R}\}$, $i = 1, \dots, q$, *are mutually independent standard Wiener processes.*

3.4. Some non-asymptotic results for i.i.d. normal innovations

Following [19], we derive some non-asymptotic results in the case where the innovations are i.i.d. Gaussian with known variance σ^2. Furthermore, we concentrate on the classical setting, where the number of changes, change points and magnitude of changes are non-random but can depend on the sample size n.

To keep the arguments simpler, we require the distance between any two change points to be larger than $2G$. We then apply the procedure from Section 3.1 with $\hat{\tau}_{k,n}^2 = \sigma^2$ and the critical value c_n instead of $D_n(G; \alpha_n)$. The critical value c_n is essentially of the same order as $D_n(G; \alpha_n)$ in the previous sections but we use a different notation as the asymptotic interpretation of $D_n(G; \alpha_n)$ no longer makes sense here.

Theorem 3.4. *Let* $\{\varepsilon_i\}$ *be i.i.d.* $N(0, \sigma^2)$, $d_j = d_{j,n}$, $k_j = k_{j,n}$, q_n *deterministic with*

$$d_j \ge \delta_n > 0, \qquad \min_{0 \le j \le q_n} |k_{j+1} - k_j| > 2G$$

and $\hat{\tau}_{k,n} = \sigma$. *Furthermore, let the critical value* c_n *fulfill*

$$c_n > \sqrt{4 \log n}, \qquad \sqrt{2} c_n \le \eta \delta_n \sqrt{G} - \sqrt{8 \log n}. \tag{3.5}$$

(a) *Then, it holds for some constant* $C > 0$

$$P\Big(\hat{q}_n = q_n, \max_{1 \le j \le q_n} |\hat{k}_j - k_j| < G\Big) \ge 1 - \frac{C}{T}.$$

(b) *It holds for any* $1 \le j \le q_n$ *as well as* $1 \le \xi_n \le G$ *and any* $\gamma > 2$

$$\text{(i)} \quad P\big(|\hat{k}_j 1_{\{j \le \hat{q}_n\}} - k_j| > \xi_n\big) \le \widetilde{C}\delta_n^{-\gamma}\xi_n^{-\frac{\gamma}{2}} - \frac{C}{T},$$

$$\text{(ii)} \quad P\Big(\max_{1 \le j \le q_n} |\hat{k}_j 1_{\{j \le \hat{q}_n\}} - k_j| > \xi_n\Big) \le \widetilde{C} q_n \delta_n^{-\gamma}\xi_n^{-\frac{\gamma}{2}} - \frac{C}{T}$$

for some constant $\widetilde{C} > 0$ *and* C *as in* (a).

There are several interesting observations about Theorem 3.4. First of all, (3.5) is closely related to Assumption A.1(a) (for fixed $\eta > 0$) as well as Assumption 3.3 in [19] for the wild binary segmentation procedure since δ_n in our notation is a lower bound for the magnitude of changes and $2G$ is a lower bound for the distance between change points. Similarly, to their result, we do not require that the number of changes is bounded (except for the assumption on the distance between change points). Secondly, if the number of changes is bounded, that is, $q_n \le q < \infty$, and the magnitude is bounded from below, that is, $\delta_n \ge \delta > 0$, then the assertion in (b)(ii) yields

$$\max_{1 \le j \le q_n} |\hat{k}_j 1_{\{j \le \hat{q}_n\}} - k_j| = O_P(1),$$

which is slightly stronger than the statements in Theorems 3.1 and 3.2 in [19] for the binary and wild binary segmentation algorithm.

3.5. Choice of bandwidth and alternative evaluations of MOSUM graphs

In some extreme cases with small bandwidth G and very close changes as in Figures 4 or 5 below the condition $w_j - v_j \ge \eta G$ in (3.1) (eta-check) can be too restrictive. In order to get asymptotic consistency of the change point estimators, it has to be replaced by another condition avoiding the double estimation of the same change point. One alternative solution checks for each significant point $\tilde{k}_j$ whether it is also a local maximum in the sense of

$$\frac{|T_{\tilde{k}_j,n}(G)|}{\hat{\tau}_{\tilde{k}_j,n}} = \max_{|\tilde{k}_j - k| < \lfloor cG \rfloor} \frac{|T_{k,n}(G)|}{\hat{\tau}_{k,n}} \qquad \text{(maximum-check)},$$

where $c > 0$ in a similar spirit as [46]. Significant local maxima in that sense are included in the list of estimated change points. Similar arguments as in the proof of Theorem 3.4 can be applied to show that there is only one significant local maximum in each environment of a true change point with probability approaching one. In practice, this method also detects changes that are more than cG but less than $2G$ apart, where the graph may sometimes not fall beneath the asymptotic critical value between changes (see Figure 5 below).

The main problem with the MOSUM statistic as investigated in this paper is that its performance depends crucially on the choice of the bandwidth G. Theoretically, the bandwidth should be chosen as large as possible with the restriction that there should not be more than one change point in any window of size $2G$ (see also Assumption A.2), that is, G should be half the minimal distance between two change points. The simulations below suggest that a somewhat larger window results in a good performance as well.

There are two main problems with the bandwidth: First, the distance between change points is not known. Secondly, while change point tests can principally detect large changes even between short stationary stretches, longer stationary stretches (at least to one side) are required to detect smaller changes. For the above MOSUM procedure, this means that performance will suffer in data sets, where both large changes surrounded by small stationary stretches and small changes surrounded by large stationary stretches are present, when only one bandwidth is used.

One solution to this problem is to use the method merely as a diagnostic tool that can be run with several bandwidths. Large changes surrounded by either large or small stationary stretches will be detected with a small bandwidth, while smaller changes surrounded by longer stationary stretches will be detected by larger bandwidths. This can be visualized using plots similar to Figures 3, 4 and 5.

Another solution is an appropriate merging of change point candidates obtained from different bandwidths. This is the idea behind the multiscale method proposed by [33] which is based on their MOSUM statistics for detection of changes in point processes: All change point candidates obtained from the MOSUM statistics based on the smallest bandwidth are included, then one proceeds recursively with the next largest bandwidth but only includes candidates if there is

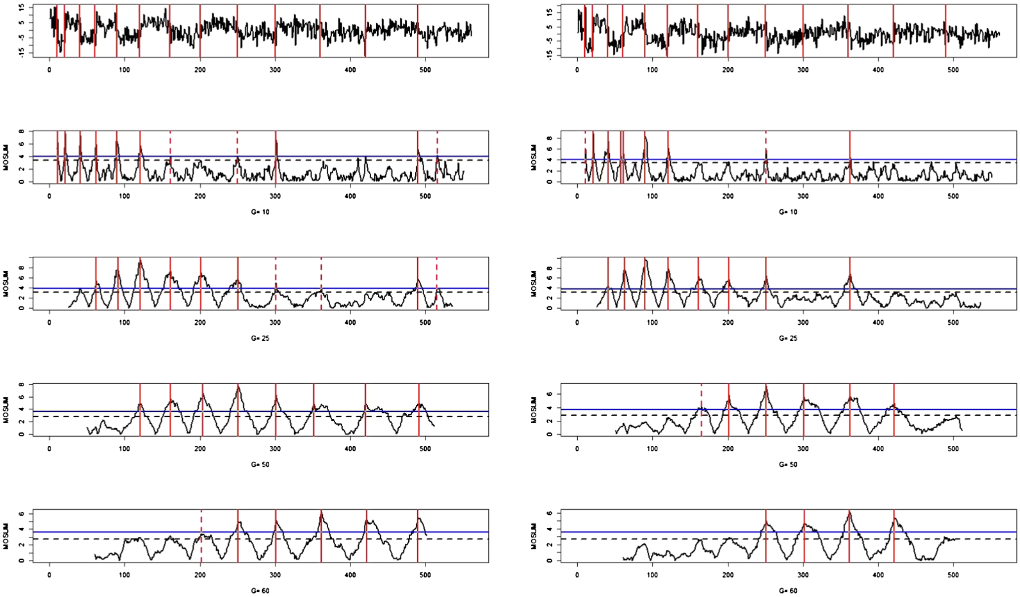

Figure 3. Two sample path and performance of $\hat{\sigma}_{k,n}^{-1} T_{k,n}(G)$ with bandwidths $G = 10$, 25, 50, 60 and $\eta = 0.15$ for “Mix”-Signal, where critical values and detected change points for $\alpha = 0.1$ are indicated by vertical solid lines and for $\alpha = 0.5$ by vertical dotted lines. The horizontal lines give the asymptotic critical values for $\alpha = 0.1$ (solid) and $\alpha = 0.5$ (dotted).

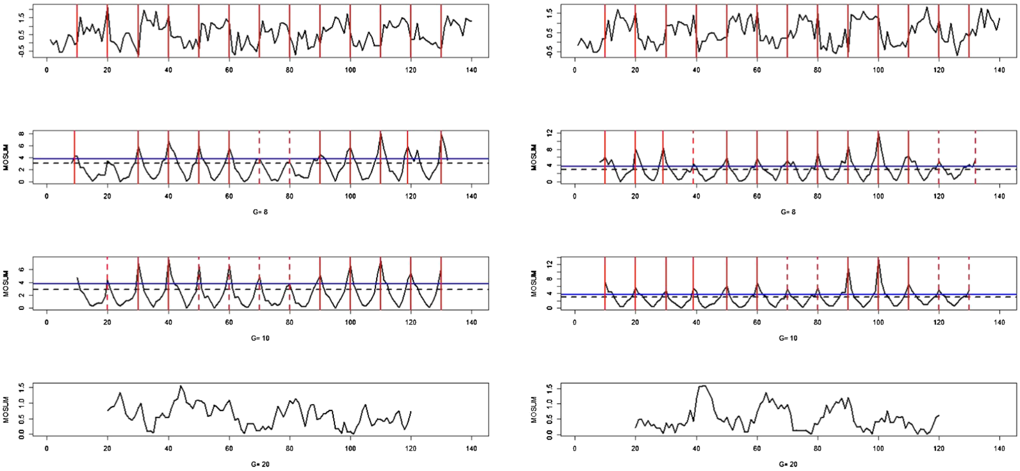

Figure 4. Two sample path and performance of $\hat{\sigma}_{k,n}^{-1}T_{k,n}(G)$ with bandwidths $G = 8$, 10, 20 and $\eta = 0.15$ for "Teeth"-Signal, where critical values and detected change points for $\alpha = 0.1$ are indicated by vertical solid lines and for $\alpha = 0.5$ by vertical dotted lines. The horizontal lines give the asymptotic critical values for $\alpha = 0.1$ (solid) and $\alpha = 0.5$ (dotted).

no previously detected change point in a suitable environment of the candidate. We will call this procedure based on the MOSUM method in this paper "Merged MOSUM (bandwidth)". One major drawback is that it is not clear how to generalize it to situations when asymmetric

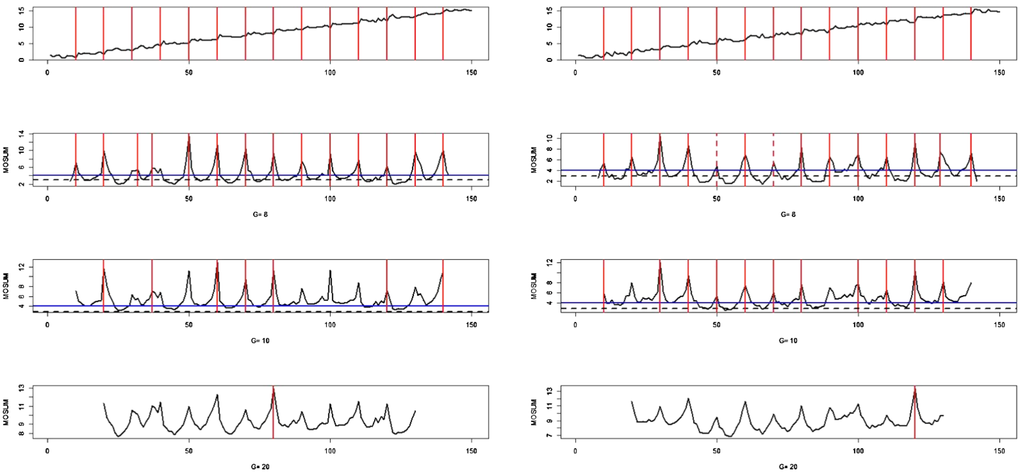

Figure 5. Two sample path and performance of $\hat{\sigma}_{k,n}^{-1}T_{k,n}(G)$ with bandwidths $G = 8$, 10, 20 and $\eta = 0.15$ for "Stairs"-Signal, where critical values and detected change points for $\alpha = 0.1$ are indicated by vertical solid lines and for $\alpha = 0.5$ by vertical dotted lines. The horizontal lines give the asymptotic critical values for $\alpha = 0.1$ (solid) and $\alpha = 0.5$ (dotted).

bandwidths, that is, different bandwidth lengths to the right and left, are used. Those would be useful to detect small changes with only a long stationary stretch to one side.

An alternative merging method (called "Merged MOSUM (p-value)") proceeds similarly to [19] where change point candidates are included according to increasing p-values unless a change point has already been detected in a suitable environment of the candidate. This algorithm can easily be adapted to allow for asymmetric bandwidths. However, care has to be taken as the p-value only corrects for multiple testing within each set of bandwidth but not across different sets of bandwidths.

In the simulations below, we implement both merging algorithms above based on symmetric bandwidths from a given list $G = G_1, \dots, G_k$, where in both cases the maximum-check with $c = 2/3$ is used instead of the eta-check. The algorithms are based on level $\alpha = 0.1$ resp. p-values smaller than 0.1. When merging, we check whether the newly detected change point is at least $c = 2/3$ of its bandwidth away from the previously accepted change point candidates. In the future some more finetuning of the method is needed and will be implemented in the R-package "Mosum".

Yet another alternative to be investigated in the future is the use of an information criterion based on possible change points obtained from one or several MOSUM statistics similarly to [19], who suggests to use an information criterion effectively as a stopping rule for the wild binary segmentation method. This should take care of a possible overestimation if one includes very small bandwidths in situations where changes are relatively far apart. It should also deal with the multiple testing issue present when multiple bandwidths are used. A somewhat different but related approach is taken by [46], who use the local maxima of a MOSUM statistic with a single bandwidth as possible change point candidates in order to reduce computational complexity of a full information criterion based approach.

4. Simulations

In this section, we illustrate the performance of the above change point estimators based on several toy data sets. Nonparametric estimators for the long-run variance are relatively imprecise even for larger sample sizes as this is a difficult statistical estimation problem, so that we use i.i.d. errors in these simulations. Furthermore, all time series are created using the deterministic change point setting because this allows us to play with the locations of the change points and see how they influence the estimators. All simulations are based on 1000 repetitions.

We will now show how the procedure works using some examples from [19] with i.i.d. normal errors. We run the simulations with the other competing procedures that have also been considered in [19] in precisely the same way, where we use the suggested sSIC-method for wild binary segmentation. Precisely, the methods are: Segmentor from the Segmentor3IsBack-R-Package, CumSeg as described in [34], SMUCE as described in [18] as well as PELT as described in [23]. We exclude the Bai and Perron procedure [4] due to its extensive computation time, although some preliminary results show that it works very well if the parameter h is chosen sufficiently small (which however increases computation time even further). Furthermore, it should be mentioned that PELT [23] yields much better results after the new version has been installed. For the MOSUM method we include the two described merging algorithms with the same bandwidths choices, where results are very similar. The change point estimates in the plots indicated by vertical lines are obtained using the MOSUM method with a single bandwidth choice, where the eta-criterion with $\eta = 0.15$ was used.

We will first look at the "Mix" signal, which has change points at 10, 20, 40, 60, 100, 120, 160, 200, 250, 300, 360, 420, 490 with means 74, -7, 6, -6, 5, -5, ... Because the signal includes large changes surrounded by small stationary stretches as well as small changes surrounded by large stationary stretches no MOSUM procedure with a single bandwidth can detect all changes. Figure 3 illustrates how different bandwidths detect different changes. Frequent change points to be missed are the ones to the very left respectively very right of the time series. As already indicated by [19], the signal is difficult to detect for all methods. Table 1 gives the results for the two merged MOSUM procedures and the WBS method only because the other competing procedures detect the correct number of change points in less than 19% of the cases. The L_1-error is the sum of the absolute differences between true change point and corresponding estimator hence measures the precision of the estimated locations.

Next, we look at a "Teeth" signal with a mean change every 10th observation switching between 0 and 1 with i.i.d. normal errors with a standard deviation of 0.4. The changes are relatively close so that only small bandwidths have a chance of detecting them. Figure 4 shows the results for different bandwidth choices and $\alpha = 0.1$ as well as $\alpha = 0.5$. The bandwidths $G = 8$ as well as $G = 10$ can well detect change points, while $G = 20$ results in oversmoothing with insignificant results. In this example, one can clearly see the problems with the eta-criterion (3.1) which have already been described in Section 3.5. This is the reason why $\alpha = 0.1$ finds so little change points although most peaks do cross the line (but only for one data point). A visual inspection, on the other hand, can clearly distinguish between those scenarios and would usually find all 10 change points. The maximum-check is much closer to this visual inspection and therefore performs better in this example. Table 2 gives the results for the merged MOSUM and WBS only as PELT, CumSeg and SMUCE detect the correct number of changes in less than 10% of the cases, and Segmentor detects it in only 52% of the cases. While WBS is slightly better in picking the correct

Table 1. Number of correctly estimated change points and L_1-error for WBS and merged MOSUM-procedure (with bandwidths $G = \{10, 25, 50, 60\}$) for the mix signal based on those realizations where both methods detect the correct number of change points

$\hat{q}_n - q$	≤ 3	$= 2$	$= 1$	$= 0$	$= 1$	$= 2$	≥ 3
Merged MOSUM (bandwidth)	0.006	0.103	0.389	0.418	0.073	0.010	0.001
Merged MOSUM (p-value)	0.006	0.091	0.353	0.432	0.098	0.018	0.002
WBS	0.068	0.301	0.268	0.335	0.022	0.004	0.002

L_1-error	Median (mean)	MAD (SD)
Merged MOSUM (bandwidth)	28 (38.84)	10.38 (47.73)
Merged MOSUM (p-value)	27 (36.64)	8.9 (46.97)
WBS	29 (36.37)	13.34 (39.31)

Table 2. Number of correctly estimated change points and L_1-error for WBS and merged MOSUM-procedure (with bandwidths $G = \{10, 25, 50, 60\}$) for the "Teeth" signal based on those realizations where both methods detect the correct number of change points

$\hat{q}_n - q$	≤ 3	$= 2$	$= 1$	$= 0$	$= 1$	$= 2$	≥ 3
Merged MOSUM (bandwidth)	0.016	0.075	0.193	0.716	0	0	0
Merged MOSUM (p-value)	0.016	0.075	0.193	0.716	0	0	0
WBS	0.059	0.027	0.016	0.735	0.129	0.020	0.014

L_1-error	Median (mean)	MAD (SD)
Merged MOSUM (bandwidth)	0 (0.55)	0 (0.93)
Merged MOSUM (p-value)	0 (0.55)	0 (0.93)
WBS	4 (4.48)	2.97 (3.12)

number of change points the location of the change point estimators are much more precise for the merged MOSUM algorithms.

Finally, we look at the "Stairs" signal in [19] with change points at every 10th observation, mean values of $1, 2, 3, \ldots$ and i.i.d. normal errors with a standard deviation of 0.3. Again a visual inspection of Figure 5 gives a clear picture, where the change points are. However, there are several problems: The eta-check can again result in missing change points, so that the maximum-check yields better results. Furthermore, it happens (particularly for $G = 10$) that the graph remains above the critical value at all times. By construction the original algorithm cannot deal with this, although clearly significant points further apart than $2G$ cannot be caused by the same change point, the implemented maximum-check on the other hand does not require the curve to fall beneath the critical level in between estimated change points resulting in better results. As already mentioned the maximum-check is closer to the visual check in spirit. Table 3 gives the results for the merged MOSUM procedures with bandwidths $G = 8, 10, 20, 30, 50$ in addition to the results of PELT, cumseg and WBS. SMUCE is excluded because it only detects 7.3% of the changes. In this example detection rate and accuracy of the estimators are very good for the merged MOSUM procedures.

The simulations for the fms and blocks signal in [19] show that both merging algorithms tend to overestimate the number of change points with the set of bandwidths we used. A possible solution could be the use of an information criterion as final pruning step that has been discussed in Section 3.5.

The following additional simulations can be found in the supplementary material [16]:

- Impact of the moving variance estimator,
- adaptations for mean changes accompanied by a variance change,
- connection between bandwidth, distance between changes and magnitude of changes.

Table 3. Number of correctly estimated change points and L_1-error for WBS and merged MOSUM-procedure (with bandwidths $G = \{8, 10, 20, 30, 50\}$) for the "Stairs" signal based on those realizations where all of those methods detect the correct number of change points

$\hat{q}_n - q$	≤ 3	$= 2$	$= 1$	$= 0$	$= 1$	$= 2$	≥ 3
CumSeg	0.014	0.02	0.093	0.787	0.086	0	0
Merged MOSUM (bandwidth)	0	0.001	0.027	0.972	0	0	0
Merged MOSUM (p-value)	0	0.001	0.028	0.971	0	0	0
PELT	0.002	0.007	0.072	0.916	0.003	0	0
WBS	0	0	0.001	0.603	0.299	0.068	0.029

L_1-error	Median (mean)	MAD (SD)
CumSeg	7 (7.56)	2.97 (2.9)
Merged MOSUM (bandwidth)	2 (1.881)	1.48 (1.52)
Merged MOSUM (p-value)	1 (1.03)	1.48 (1.26)
WBS	2 (2.18)	1.48 (1.75)
PELT	1 (1.68)	1.48 (1.55)

Conclusions: There are several advantages of the MOSUM procedure: First, it provides some meaningful graphs for visual inspection. Secondly, p-values are attached to change point estimators with the usual interpretation at least for a given bandwidth. Third, the correctly discovered MOSUM change point estimators are very precise. Finally, we have demonstrated the potential of merged MOSUM algorithms, where theoretic and practical details are left to be investigated in future work.

5. Proofs

In this section, we prove the results of the previous sections. More detailed proofs can be found in [35].

5.1. Proofs of Section 2

Proof of Theorem 2.1. By the invariance principle in Assumption A.1(b) and (2.3), we get

$$\max_{G \le k \le n-G} \frac{1}{\sqrt{2G\tau^2}} \left| \sum_{i=k+1}^{k+G} \varepsilon_i - \big(W(k+G) - W(k)\big) \right| = O_P\left(\frac{n^{1/(2+\nu)}}{\sqrt{2G}} \right) = o_P\big(a(n/G)^{-1}\big).$$

Consequently, we can replace X_i/τ in the statistic by i.i.d. standard normal random variables $\tilde{\varepsilon}_i$ without changing the asymptotics. The assertion then follows from Theorem 2.1 in [22]. This implies

$$\max_{G\le k\le n-G}\left|T_{k,n}(G)\right| = O_P\left(\sqrt{\log(n/G)}\right) \tag{5.1}$$

showing that we can replace τ by $\hat{\tau}_{k,n}$ without changing the asymptotics since by assumption (2.5) it holds

$$\max_{G\le k\le n-G}\left|\frac{1}{\hat{\tau}_{k,n}}-\frac{1}{\tau}\right| = o_P\left(\left(\log(n/G)\right)^{-1}\right).$$ □

Part (b) of the following lemma is essentially a corollary of Theorem 2.1, showing that the statistic away from change points behaves exactly as under the null hypothesis, while (a) immediately implies consistency of the test under alternatives. Combined they will be the key to proving consistency of the estimator for the number of changes in the next section.

Lemma 5.1. *Let the assumptions of Theorem 2.2 hold.*

(a) *If*

$$\max_{G\le k\le n-G}\hat{\tau}_{k,n}^2 = o_P\left(\min_{j=1,\ldots,q_n} d_j^2 \frac{\log(n/G)G}{c_{\alpha_n}^2}\right) \tag{5.2}$$

for some sequence $\{\alpha_n\}$ *with* $0<\alpha_n<1$, *then for any* $\varepsilon>0$

$$P\left(\min_{\substack{0\le|k-k_j|<(1-\varepsilon)G\\ j=1,\ldots,q_n}}\frac{|T_{k,n}(G)|}{\hat{\tau}_{k,n}} < D_n(G;\alpha_n)\right) \longrightarrow 0,$$

where c_{α_n}, $D_n(G;\alpha_n)$ *as in* (2.6).

(b) *If (as in Theorem* 3.1)

$$\max_{|k-k_j|\ge G, j=1,\ldots,q_n}\left|\hat{\tau}_{k,n}^2-\tau^2\right| = o_P\left(\left(\log(n/G)\right)^{-1}\right),$$

then it holds uniformly in α

$$P\left(\max_{\substack{|k-k_j|\ge G\\ j=1,\ldots,q_n}}\frac{|T_{k,n}(G)|}{\hat{\tau}_{k,n}} \ge D_n(G;\alpha)\right) \le \alpha.$$

Proof. To obtain (a), note that by assumption it is sufficient to prove the assertion if additionally $\min_{0\le j\le q_n}|k_{j+1}-k_j| > 2G$ holds. In this case some calculations (for details, we refer to [35], proof of Theorem 6.1) give for $0\le|k-k_j|\le(1-\varepsilon)G$

$$T_{k,n}(G;X_1,\ldots,X_n) = T_{k,n}(G;\varepsilon_1,\ldots,\varepsilon_n) + \frac{d_j}{\sqrt{2G}}\left(G-|k-k_j|\right).$$

Hence, by (5.1)

$$\begin{aligned}\min_{0\le|k-k_j|<(1-\varepsilon)G}\bigl|T_{k,n}(G)\bigr| &= \min_{0\le|k-k_j|<(1-\varepsilon)G}\left|T_{k,n}(G;\varepsilon_1,\dots,\varepsilon_n)+\frac{d_j(G-|k-k_j|)}{\sqrt{2G}}\right|\\ &\ge \frac{\min_{1\le j\le q_n}|d_j|}{\sqrt{2}}\varepsilon\sqrt{G}-\max_{G\le k\le n-G}\bigl|T_{k,n}(G;\varepsilon_1,\dots,\varepsilon_n)\bigr|\\ &= \frac{\min_{1\le j\le q_n}|d_j|}{\sqrt{2}}\varepsilon\sqrt{G}+O_P\bigl(\sqrt{\log(n/G)}\bigr).\end{aligned}$$

We conclude

$$\begin{aligned}&P\biggl(\min_{0\le|k-k_j|<(1-\varepsilon)G}\frac{|T_{k,n}(G)|}{\hat\tau_{k,n}}<D_n(G,\alpha_n),\ \min_{1\le j\le q_n}|k_{j+1}-k_j|>2G,\ \min_{G\le k\le n-G}\hat\tau_{k,n}>0\biggr)\\ &\le P\biggl(\Bigl(\min_{1\le j\le q_n}|d_j|\Bigr)^{-1}G^{-1/2}\biggl(\max_{G\le k\le n-G}\hat\tau_{k,n}\frac{c_{\alpha_n}+b(n/G)}{a(n/G)}+O_P\bigl(\sqrt{\log(n/G)}\bigr)\biggr)>\frac{\varepsilon}{\sqrt{2}}\biggr)\\ &\to 0,\end{aligned}$$

proving (a). Assertion (b) follows immediately from Theorem 2.1 on noting that on each segment away from any k_j, the MOSUM statistic as well as the variance estimator behave exactly as under H_0 due to Assumption A.1(a), where the convergence is uniformly in α due to the continuity of the Gumbel limit distribution. □

Proof of Theorem 2.2. It follows immediately from Lemma 5.1(a) on noting that Assumption A.4(b) implies (5.2) for fixed α resp. c_α. □

Proof of Theorem 2.3. First, we obtain

$$\begin{aligned}&P\Biggl(\max_{0\le k\le n-G}\Biggl|\frac{1}{G}\sum_{h=0}^{\Lambda_n}\omega(h/\Lambda_n)\sum_{i=k+1}^{k+G-h}\bigl(\varepsilon_i\varepsilon_{i+h}-\gamma(h)\bigr)\Biggr|>\epsilon\Biggr)\\ &\le\sum_{k=0}^{n-G}P\Biggl(\Biggl|\sum_{h=0}^{\Lambda_n}\omega(h/\Lambda_n)\sum_{i=k+1}^{k+G-h}\bigl(\varepsilon_i\varepsilon_{i+h}-\gamma(h)\bigr)\Biggr|>G\epsilon\Biggr)\\ &\le\frac{n}{G^2\epsilon^2}\sum_{h=0}^{\Lambda_n}\sum_{b=0}^{\Lambda_n}\omega(h/\Lambda_n)\omega(b/\Lambda_n)\sum_{i=1}^{G-h}\sum_{j=1}^{G-b}E\bigl(\bigl(\varepsilon_i\varepsilon_{i+h}-\gamma(h)\bigr)\bigl(\varepsilon_j\varepsilon_{j+b}-\gamma(b)\bigr)\bigr)\\ &\le\frac{n}{G^2\epsilon^2}\sum_{h=0}^{\Lambda_n}\sum_{b=0}^{\Lambda_n}\sum_{i=1}^{G-h}\sum_{j=1}^{G-b}\bigl|\nu(h,j-i,j-i+b)\bigr|\le C\frac{n\Lambda_n^2}{G^2\epsilon^2}\end{aligned}$$

by assumption (2.9). Consequently,

$$\max_{G\le k\le n-G}\left|\frac{1}{G}\sum_{h=0}^{\Lambda_n}\omega(h/\Lambda_n)\sum_{i=k-G+1}^{k-h}\big(\varepsilon_i\varepsilon_{i+h}-\gamma(h)\big)\right|=O_P\left(\frac{n^{1/2}\Lambda_n}{G}\right). \tag{5.3}$$

From Assumption A.1(b) and a Hájék–Rényi-inequality as in Lemma 3.1 for the Wiener process we get

$$\max_{1\le k\le n}\left|\sum_{i=1}^{k}\varepsilon_i\right|=O_P(\sqrt{n}). \tag{5.4}$$

Since

$$\max_{G\le k\le n-G}\left|\hat\sigma^2_{k,n}-\sigma^2\right|\le\max_{0\le k\le n-G}\frac{1}{G}\left|\sum_{i=k+1}^{k+G}\big(\varepsilon_i^2-\sigma^2\big)\right|+\max_{0\le k\le n-G}\left|\frac{1}{G}\sum_{i=k+1}^{k+G}\varepsilon_i\right|^2$$

assertion (a) follows from (5.3) (with $\Lambda_n<1$) and (5.4).

Similarly,

$$\begin{aligned}
&\max_{G\le k\le n-G}\left|\hat\tau^2_{k,n}-\tau^2\right|\\
&=\max_{G\le k\le n-G}\left|\hat\sigma^2_{k,n}+2\sum_{h=1}^{\Lambda_n}\omega(h/\Lambda_n)\hat\gamma_k(h)-\sigma^2-2\sum_{h>0}\gamma(h)\right|\\
&\le\max_{G\le k\le n-G}\left|\hat\sigma^2_{k,n}-\sigma^2\right|\\
&\quad+\max_{G\le k\le n-G}\left|\sum_{h=1}^{\Lambda_n}\omega(h/\Lambda_n)\frac{1}{G}\sum_{i=k-G+1}^{k-h}(\varepsilon_i-\bar\varepsilon_{k-G+1,k})(\varepsilon_{i+h}-\bar\varepsilon_{k-G+1,k})-\sum_{h>0}\gamma(h)\right|\\
&\quad+\max_{G\le k\le n-G}\left|\sum_{h=1}^{\Lambda_n}\omega(h/\Lambda_n)\frac{1}{G}\sum_{i=k+1}^{k+G-h}(\varepsilon_i-\bar\varepsilon_{k+1,k+G})(\varepsilon_{i+h}-\bar\varepsilon_{k+1,k+G})-\sum_{h>0}\gamma(h)\right|\\
&\le O_P\left(\frac{\Lambda_n n^{1/2}}{G}\right)+\sum_{h\in\mathbb{Z}}\left|w(h/\Lambda_n)-1\right|\left|\gamma(h)\right|,
\end{aligned}$$

where we made repeated use of (5.4). This completes the proof of (b). □

Proof of Theorem 2.4. By assumption, it holds

$$X_i-\overline{X}_{k-G+1,k}=\varepsilon_i-\bar\varepsilon_{k-G+1,k}+O_P(1),\qquad i=k-G+1,\dots,k, \tag{5.5}$$

where the rate is uniform in i,k,G. Hence by $(a+b)^2\le 2a^2+2b^2$, we get

$$\sup_{G\le k\le n-G}\hat\sigma^2_{k,n}\le\sup_{G\le k\le n-G}\frac{1}{G}\left(\sum_{i=k-G+1}^{k}(\varepsilon_i-\bar\varepsilon_{k-G+1,k})^2+\sum_{i=k+1}^{k+G}(\varepsilon_i-\bar\varepsilon_{k+1,k+G})^2\right)+O_P(1),$$

implying (a) by Theorem 2.3. Similarly, by (5.5) and (5.4), we get

$$\begin{aligned}
&\frac{1}{G}\sum_{i=k+1}^{k+G-h}(X_i-\overline{X}_{k+1,k+G})(X_{i+h}-\overline{X}_{k+1,k+G})\\
&=\frac{1}{G}\sum_{i=k+1}^{k+G-h}\big(\varepsilon_i-\bar{\varepsilon}_{k+1,k+G}+O_P(1)\big)\big(\varepsilon_{i+h}-\bar{\varepsilon}_{k+1,k+G}+O_P(1)\big)\\
&=\frac{1}{G}\sum_{i=k+1}^{k+G-h}(\varepsilon_i-\bar{\varepsilon}_{k+1,k+G})(\varepsilon_{i+h}-\bar{\varepsilon}_{k+1,k+G})+O_P(1)+O_P\left(\frac{\sqrt{n}}{G}\right),
\end{aligned}$$

where the rates are uniform in k, h, G. With Theorem 2.3, we get

$$\max_{G\le k\le n-G}\hat{\tau}_{k,n}^2=O_P\left(\frac{\Lambda_n n^{1/2}}{G}\right)+O_P(\Lambda_n)=O_P(\Lambda_n)$$

concluding (b). □

5.2. Proofs of Section 3

Proof of Theorem 3.1. First, note that Assumptions A.4(b) and (3.4) imply (5.2) so that by Lemma 5.1, it holds $P(S_n)\to 1$ with

$$\begin{aligned}
S_n=\Bigg\{&\max_{|k-k_j|\ge G,\, j=1,\dots,q_n}\frac{|T_{k,n}(G)|}{\hat{\tau}_{k,n}}<D_n(G,\alpha_n),\\
&\min_{0\le|k-k_j|<(1-\eta)G,\, j=1,\dots,q_n}\frac{|T_{k,n}(G)|}{\hat{\tau}_{k,n}}\ge D_n(G,\alpha_n)\Bigg\}.
\end{aligned}$$

Since

$$S_n\subset\{\hat{q}_n=q_n\},\qquad S_n\subset\left\{\max_{1\le j\le q_n}|\hat{k}_j 1_{\{j\le\hat{q}_n\}}-k_j|\le G\right\},$$

the assertion follows. □

Proof of Lemma 3.1. By the triangle inequality and the monotony of the $\{c_k\}$, we get

$$\max_{\ell\le k\le u}c_k\left|\sum_{j=m+1}^{m+k}\varepsilon_j\right|\le c_\ell\left|\sum_{j=m+1}^{m+\ell}\varepsilon_j\right|+\max_{\ell<k\le u}c_k\left|\sum_{j=m+\ell+1}^{m+k}\varepsilon_j\right|,$$

hence

$$P\left(\max_{\ell\le k\le u}c_k\left|\sum_{j=m+1}^{m+k}\varepsilon_j\right|>\delta\right)\le P\left(c_\ell\left|\sum_{j=m+1}^{m+\ell}\varepsilon_j\right|>\delta/2\right)+P\left(\max_{\ell<k\le u}c_k\left|\sum_{j=m+\ell+1}^{m+k}\varepsilon_j\right|>\delta/2\right).$$

An index shift in connection with the Chebyshev inequality and Assumption A.5 yields (a) by Theorem B.3 in [24]. Analogous arguments show (b) on noting that Assumption A.1(c) also holds for the process in reversed time $\{\varepsilon_{-t}\}$. □

The following lemma is needed in the proofs of Theorems 3.2 as well as 3.3.

Lemma 5.2. *Let the errors fulfill Assumptions* A.1(a) *and* A.1(c). *Then it holds for any* $\beta > 0$, $\xi > 0$ *on* $2G \le k_j \le n - 2G$:

(a) $P(\max_{k_j-G\le k\le k_j-\xi} \frac{|T_{k_j,n}(G;\varepsilon_1,\dots,\varepsilon_n)-T_{k,n}(G;\varepsilon_1,\dots,\varepsilon_n)|}{k_j-k} > \beta) = O((\beta^2 G\xi)^{-\gamma/2})$,

(b) $P(\max_{k_j-u\le k\le k_j} |T_{k_j,n}(G;\varepsilon_1,\dots,\varepsilon_n) - T_{k,n}(G;\varepsilon_1,\dots,\varepsilon_n)| > \beta) = O(\beta^{-\gamma}(\frac{u}{G})^{\gamma/2})$,

(c) $P(\max_{k_j-G\le k\le k_j-\xi} |T_{k_j,n}(G;\varepsilon_1,\dots,\varepsilon_n) + T_{k,n}(G;\varepsilon_1,\dots,\varepsilon_n)| > \beta) = O(\beta^{-\gamma})$,

where the constants only depend on $\widetilde{C}$ *and* γ.

Proof of Lemma 5.2. For $G \le k_j - G \le k \le k_j - \xi$ some straightforward calculations (confer [35] (6.13)) give

$$\begin{aligned} &T_{k_j,n}(G;\varepsilon_1,\dots,\varepsilon_n) - T_{k,n}(G;\varepsilon_1,\dots,\varepsilon_n) \\ &\quad = \frac{1}{\sqrt{2G}}\left(\sum_{i=k+G+1}^{k_j+G}\varepsilon_i + \sum_{i=k-G+1}^{k_j-G}\varepsilon_i - 2\sum_{i=k+1}^{k_j}\varepsilon_i\right). \end{aligned} \tag{5.6}$$

Hence,

$$\begin{aligned} &P\left(\max_{k_j-G\le k\le k_j-\xi} \frac{|T_{k_j,n}(G;\varepsilon_1,\dots,\varepsilon_n) - T_{k,n}(G;\varepsilon_1,\dots,\varepsilon_n)|}{k_j-k} > \beta\right) \\ &\le P\left(\max_{k_j-G\le k\le k_j-\xi} \frac{|\sum_{i=k+G+1}^{k_j+G}\varepsilon_i|}{k_j-k} > \frac{\beta\sqrt{2G}}{3}\right) \\ &\quad + P\left(\max_{k_j-G\le k\le k_j-\xi} \frac{|\sum_{i=k-G+1}^{k_j-G}\varepsilon_i|}{k_j-k} > \frac{\beta\sqrt{2G}}{3}\right) \\ &\quad + P\left(\max_{k_j-G\le k\le k_j-\xi} \frac{|\sum_{i=k+1}^{k_j}\varepsilon_i|}{k_j-k} > \frac{\beta\sqrt{2G}}{3}\right). \end{aligned}$$

By Lemma 3.1 and the independence of k_j and $\varepsilon_1,\dots,\varepsilon_n$ it follows for the first summand, where the others can be dealt with analogously,

$$\begin{aligned} P\left(\max_{k_j-G\le k\le k_j-\xi}\left|\frac{\sum_{i=k+G+1}^{k_j+G}\varepsilon_i}{k_j-k}\right| > \frac{\beta\sqrt{2G}}{3}\right) &\le \widetilde{C}\beta^{-\gamma}G^{-\gamma/2}\left(\xi^{-\gamma/2} + \sum_{k=\xi+1}^{G} k^{-\gamma/2-1}\right) \\ &= O\big((\beta^2 G\xi)^{-\gamma/2}\big), \end{aligned}$$

where $O(1)$ only depends on $\widetilde{C}$ and γ. The proof of (b) is analogous (with $c_k = 1$ in Lemma 3.1). Since

$$\begin{aligned}&T_{k_j,n}(G;\varepsilon_1,\dots,\varepsilon_n) + T_{k,n}(G;\varepsilon_1,\dots,\varepsilon_n)\\ &\quad = \big(T_{k,n}(G;\varepsilon_1,\dots,\varepsilon_n) - T_{k_j,n}(G;\varepsilon_1,\dots,\varepsilon_n)\big) + 2T_{k_j,n}(G;\varepsilon_1,\dots,\varepsilon_n),\end{aligned}$$

(c) can be proven analogously, where the assertion for the first summand on the right-hand side follows from (b) and the second summand can be dealt with analogously by an application of the Chebyshev inequality in addition to Assumption A.1(c). □

Proof of Theorem 3.2. We will prove

$$\begin{aligned}P(\hat{k}_j > k_j + \xi_n) &= O(1)\delta_n^{-\gamma}\big(\xi_n^{-\frac{\gamma}{2}} + G^{-\frac{\gamma}{2}}\big) + o(1),\\ P(\hat{k}_j < k_j - \xi_n) &= O(1)\delta_n^{-\gamma}\big(\xi_n^{-\frac{\gamma}{2}} + G^{-\frac{\gamma}{2}}\big) + o(1),\end{aligned}$$

where we discuss the second assertion in detail, the first one follows analogously (where an analogous version of Lemma 5.2 is needed). Consider the set

$$\begin{aligned}M_n = &\Big\{\hat{q}_n = q_n,\ \max_{1\le j\le q_n} |\hat{k}_j - k_j| < G,\ \hat{\tau}_n > 0\Big\}\\ &\cap \Big\{\min_{j=1,\dots,q_n} |d_j| \ge \delta_n,\ \min_{j=1,\dots,q_n} |k_{j+1} - k_j| > 2G\Big\},\end{aligned}\tag{5.7}$$

which does not depend on j. Define $\tilde{w}_j := \min(w_j, k_j + G - 1)$ and $\tilde{v}_j := \max(v_j, k_j - G + 1)$, then it holds on M_n for all $j = 1,\dots,q_n$

$$\hat{k}_j = \arg\max_{\tilde{v}_j\le k\le\tilde{w}_j} \big|T_{k,n}(G)\big|.$$

Next, note that

$$\begin{aligned}\arg\max_{\tilde{v}_j\le k\le\tilde{w}_j} \big|T_{k,n}(G)\big| &= \arg\max_{\tilde{v}_j\le k\le\tilde{w}_j} V_{k,n}^{(j)}(G),\\ V_{k,n}^{(j)}(G) &= \big(T_{k,n}(G)\big)^2 - \big(T_{k_j,n}(G)\big)^2.\end{aligned}$$

Hence,

$$\begin{aligned}P\Big(\arg\max_{\tilde{v}_j\le k\le\tilde{w}_j} V_{k,n}^{(j)}(G) < k_j - \xi_n\Big) &= P\Big(\max_{\tilde{v}_j\le k\le k_j-\xi_n} V_{k,n}^{(j)}(G) \ge \max_{k_j-\xi_n\le k\le\tilde{w}_j} V_{k,n}^{(j)}(G)\Big)\\ &\le P\Big(\max_{\tilde{v}_j\le k\le k_j-\xi_n} V_{k,n}^{(j)}(G) \ge 0\Big).\end{aligned}$$

The additional term $o(1)$ in (a) represents $P(M_n^C) \to 0$ by Theorem 3.1 and assumptions. Next, for $k_j - G \le k \le k_j - \xi_n$,

$$\begin{aligned} V_{k,n}^{(j)}(G) &= \big(T_{k,n}(G)\big)^2 - \big(T_{k_j,n}(G)\big)^2 = \big(T_{k,n}(G) - T_{k_j,n}(G)\big)\big(T_{k,n}(G) + T_{k_j,n}(G)\big) \\ &= -\Big(\big(T_{k_j,n}(G;\varepsilon_1,\dots,\varepsilon_n) - T_{k,n}(G;\varepsilon_1,\dots,\varepsilon_n)\big) + (2G)^{-\frac{1}{2}}(k_j-k)d_j\Big) \\ &\qquad \times \Big(\big(T_{k_j,n}(G;\varepsilon_1,\dots,\varepsilon_n) + T_{k,n}(G;\varepsilon_1,\dots,\varepsilon_n)\big) + (2G)^{-\frac{1}{2}}(k+2G-k_j)d_j\Big) \\ &=: -\big(A_1(k,n) + D_1(k,n)\big)\big(A_2(k,n)+D_2(k,n)\big). \end{aligned} \tag{5.8}$$

Since on M_n it holds $D_1(k,n)/d_j \ge (2G)^{-1/2}(k_j-k)$ and $D_2(k,n)/d_j \ge 2^{-1/2}G^{1/2}$, we get $D_1(k,n)D_2(k,n) \ge \delta_n^2\xi_n/2 > 0$. Hence

$$\begin{aligned} &P\Big(\max_{k_j-G\le k\le k_j-\xi_n} V_{k,n}^{(j)}(G) \ge 0, M_n\Big) \\ &= P\Bigg(\max_{k_j-G\le k\le k_j-\xi_n} \\ &\qquad -D_1(k,n)D_2(k,n)\Bigg(1 + \frac{A_1(k,n)}{D_1(k,n)}\frac{A_2(k,n)}{D_2(k,n)} + \frac{A_1(k,n)}{D_1(k,n)} + \frac{A_2(k,n)}{D_2(k,n)}\Bigg) \ge 0, M_n\Bigg) \\ &\le P\Bigg(\max_{k_j-G\le k\le k_j-\xi_n}\bigg|\frac{A_1(k,n)}{D_1(k,n)}\frac{A_2(k,n)}{D_2(k,n)} + \frac{A_1(k,n)}{D_1(k,n)} + \frac{A_2(k,n)}{D_2(k,n)}\bigg| \ge 1, M_n\Bigg) \\ &\le P\Bigg(\max_{k_j-G\le k\le k_j-\xi_n}\bigg|\frac{A_1(k,n)}{D_1(k,n)}\bigg| \max_{k_j-G\le k\le k_j-\xi_n}\bigg|\frac{A_2(k,n)}{D_2(k,n)}\bigg| + \max_{k_j-G\le k\le k_j-\xi_n}\bigg|\frac{A_1(k,n)}{D_1(k,n)}\bigg| \\ &\qquad + \max_{k_j-G\le k\le k_j-\xi_n}\bigg|\frac{A_2(k,n)}{D_2(k,n)}\bigg| \ge 1, M_n\Bigg) \\ &\le P\Bigg(\max_{k_j-G\le k\le k_j-\xi_n}\bigg|\frac{A_1(k,n)}{k_j-k}\bigg| \ge \frac{\delta_n}{3\sqrt{2G}}\Bigg) + P\Bigg(\max_{k_j-G\le k\le k_j-\xi_n}\big|A_2(k,n)\big| \ge \frac{\delta_n\sqrt{G}}{3\sqrt{2}}\Bigg) \\ &\qquad + P\Bigg(\max_{k_j-G\le k\le k_j-\xi_n}\bigg|\frac{A_1(k,n)}{k_j-k}\bigg| \ge \frac{\delta_n}{\sqrt{6G}}\Bigg)P\Bigg(\max_{k_j-G\le k\le k_j-\xi_n}\big|A_2(k,n)\big| \ge \frac{\delta_n\sqrt{G}}{\sqrt{6}}\Bigg) \\ &= O(1)\delta_n^{-\gamma}\big(\xi_n^{-\frac{\gamma}{2}} + G^{-\frac{\gamma}{2}}\big), \end{aligned}$$

where the last line follows from Lemma 5.2 and $O(1)$ does not depend on j. This concludes the proof of (a) by $\xi_n \le G$. Similarly, on the set

$$\widetilde{M}_n = M_n \cap \{q_n \le \gamma_n\} \tag{5.9}$$

we get

$$P\Big(\max_{j=1,\dots,\min(\hat{q}_n,q_n)} |\hat{k}_j - k_j| > \xi_n, \widetilde{M}_n\Big) \le \sum_{j=1}^{\gamma_n} P\big(|\hat{k}_j 1_{\{j\le\min(\hat{q}_n,q_n)\}} - k_j| > \xi_n, \widetilde{M}_n\big)$$
$$= O(1)\gamma_n\delta_n^{-\gamma}\big(\xi_n^{-\frac{\gamma}{2}} + G^{-\frac{\gamma}{2}}\big),$$

where the last line follows from the proof of (a). Since $P(\widetilde{M}_n) \to 1$, the proof of (b) is complete. □

Before we can get to the proof of the corollary, we first need to refine the assertions on the variance estimator $\hat{\sigma}^2_{k,n}$ from Section 2.3:

Lemma 5.3. *Let the assumptions of Corollary 3.1 be fulfilled. Then it holds*

$$\text{(a)} \quad P\Big(\min_{0\le|k-k_j|\le G} \hat{\sigma}^2_{k,n} \ge C\Big) \to 1 \qquad \textit{for some } c > 0.$$

$$\text{(b)} \quad 2G\frac{\hat{\sigma}^2_{k,n} - \hat{\sigma}^2_{k_j,n}}{|k_j - k|} = d_j^2\frac{G - |k-k_j|}{G} + R_k(j),$$

$$\textit{where} \max_{\xi_n\le|k-k_j|\le G}\big|R_k(j)\big| = O_P\big(\xi_n^{-1/2}\big) + o_P(1).$$

Proof. Some calculations show that for $k \le k_j$

$$\hat{\sigma}^2_{k,n} = \frac{d_j^2}{2G}|k_j - k|\frac{G-|k-k_j|}{G} + \frac{1}{2G}\sum_{i=k-G+1}^{k+G}\varepsilon_i^2 - \frac{1}{2}\big(\bar{\varepsilon}^2_{k-G+1,k} + \bar{\varepsilon}^2_{k+1,k+G}\big)$$
$$- d_j\frac{1}{G}\sum_{i=k+1}^{k_j}\varepsilon_i + d_j\frac{k_j-k}{G}\bar{\varepsilon}_{k+1,k+G}.$$

From this, assertion (a) follows (symmetry arguments give the result for $k > k_j$). For (b) note that this implies (for $k_j - G \le k < k_j$)

$$2G\frac{\hat{\sigma}^2_{k,n} - \hat{\sigma}^2_{k_j,n}}{k_j - k} = d_j^2\frac{G-|k-k_j|}{G} + \frac{1}{k_j-k}\sum_{i=k-G+1}^{k_j}\big(\varepsilon_i^2 - \sigma^2\big) - \frac{1}{k_j-k}\sum_{i=k+G+1}^{k_j+G}\big(\varepsilon_i^2 - \sigma^2\big)$$
$$+ \frac{1}{k_j-k}\left(\sum_{i=k-G+1}^{k_j-G}\varepsilon_i - \sum_{i=k+1}^{k_j}\varepsilon_i\right)\big(\bar{\varepsilon}_{k-G+1,k} + \bar{\varepsilon}_{k_j-G+1,k_j}\big)$$
$$+ \frac{1}{k_j-k}\left(\sum_{i=k+1}^{k_j}\varepsilon_i - \sum_{i=k+G+1}^{k_j+G}\varepsilon_i\right)\big(\bar{\varepsilon}_{k+1,k+G} + \bar{\varepsilon}_{k_j+1,k_j+G}\big)$$
$$- 2d_j\frac{1}{k_j-k}\sum_{j=k+1}^{k_j}\varepsilon_i + 2d_j\bar{\varepsilon}_{k+1,k+G}$$

$$= d_j^2 \frac{|k-k_j|+G}{G} + R_k(j,1) + \cdots + R_k(j,6).$$

An application of the Hájek–Rényi inequality yields

$$\max_{\xi_n \le k_j - k \le G} \bigl| R_k(j,l) \bigr| = O_P\bigl(\xi_n^{-1/2}\bigr) \qquad \text{for } l = 1,2,5$$

as well as the first factor in $R_k(j,3)$ and $R_k(j,4)$. Furthermore, by another application of the Hájek–Rényi inequality (or the strong law of large numbers) we get $\max_{|k-k_j|\le G} |\bar{\varepsilon}_{k+1,k+G}| = o_P(1)$ resulting in

$$\max_{\xi_n \le k_j - k \le G} \bigl| R_k(j,l) \bigr| = o_P(1) \qquad \text{for } l = 6.$$

An analogous argument applies to the second factor in $R_k(j,3)$ as well as $R_k(j,4)$, yielding assertion (b), where the assertion for $k > k_j$ follows by symmetry arguments. □

Proof of Corollary 3.1. The proof is very close to the proof of Theorem 3.2 on replacing $V_{k,n}^{(j)}(G)$ by

$$\widetilde{V}_{k,n}^{(j)} = \frac{T_{k,n}^2}{\hat{\sigma}_k^2} - \frac{T_{k_j,n}^2}{\hat{\sigma}_{k_j}^2}$$

with the following decomposition ($\hat{\sigma}_k = \hat{\sigma}_{k,n}$)

$$\widetilde{V}_{k,n}^{(j)} = -\bigl(\widetilde{A}_1(k,n) + \widetilde{D}_1(k,n)\bigr)\bigl(\widetilde{A}_2(k,n) + \widetilde{D}_2(k,n)\bigr),$$

where

$$\begin{aligned}
\widetilde{A}_1(k,n) &= \frac{A_1(k,n)}{\hat{\sigma}_k} + \widetilde{A}_{1,2}(k,n), \\
\widetilde{A}_{1,2}(k,n) &= T_{k_j,n}(G; \varepsilon_1, \ldots, \varepsilon_n)\left(\frac{1}{\hat{\sigma}_{k_j}} - \frac{1}{\hat{\sigma}_k}\right), \\
\widetilde{A}_2(k,n) &= \frac{A_2(k,n)}{\hat{\sigma}_k} + \widetilde{A}_{1,2}(k,n), \\
\widetilde{D}_1(k,n)/d_j &= \frac{1}{\sqrt{2G}} \frac{k_j - k}{\hat{\sigma}_k} + \sqrt{\frac{G}{2}}\left(\frac{1}{\hat{\sigma}_{k_j}} - \frac{1}{\hat{\sigma}_k}\right) \\
&= \frac{D_1(k,n)}{d_j}\left(\frac{1}{\hat{\sigma}_k} + \frac{1}{\hat{\sigma}_{k_j}\hat{\sigma}_k(\hat{\sigma}_{k_j} + \hat{\sigma}_k)} G \frac{\hat{\sigma}_k^2 - \hat{\sigma}_{k_j}^2}{k_j - k}\right) =: \frac{D_1(k,n)}{d_j} \widetilde{D}_{1,2}(k,n), \\
\widetilde{D}_2(k,n)/d_j &= \frac{1}{\sqrt{2G}}\left(\frac{G}{\hat{\sigma}_{k_j}} + \frac{G + k - k_j}{\hat{\sigma}_k}\right).
\end{aligned}$$

Consider the set

$$S_n = \Big\{ \min_{\xi_n \le k_j - k \le G} \widetilde{D}_{1,2}(k,m) \ge c_1 \Big\} \cap \Big\{ \max_k \hat{\sigma}_k \le c_2 \Big\} \cap \Big\{ \min_{0 \le |k_j - k| \le G} \hat{\sigma}_k \ge c_3 \Big\}.$$

By Lemma 5.3 and Theorem 2.4 it holds $P(S_n) \to 1$ for suitable constants $c_i > 0$, $i = 1, 2, 3$, and $\xi_n \ge \xi_0$ for some $\xi_0 > 0$. Furthermore on S_n it holds (where c is a generic constant which may differ from line to line)

$$\frac{\widetilde{D}_1(k,n)}{d_j} \ge c \frac{D_1(k,n)}{d_j}, \qquad \frac{\widetilde{D}_2(k,n)}{d_j} \ge c\sqrt{G},$$
$$\frac{|A_j(k,n)|}{\hat{\sigma}_k} \le \frac{|A_j(k,n)|}{c}, \qquad j = 1, 2.$$

Finally, on S_n it holds

$$\begin{aligned} \max_{\xi_n \le k_j - k \le G} \left| \frac{\widetilde{A}_{1,2}(k,n)}{\widetilde{D}_1(k,n)} \right| &\le c \frac{1}{\sqrt{d_j^2 G}} \big| T_{k_j,n}(G; \varepsilon_1, \dots, \varepsilon_n) \big| 2G \max_{\xi_n \le k_j - k \le G} \frac{|\hat{\sigma}_k^2 - \hat{\sigma}_{k_j}^2|}{|k_j - k|} \\ &= O_P\left(\frac{1}{\sqrt{d_j^2 G}} \right) = o_P(1) \end{aligned}$$

by an application of Lemma 5.3 and the central limit theorem. Similarly,

$$\max_{\xi_n \le k_j - k \le G} \left| \frac{\widetilde{A}_{1,2}(k,n)}{\widetilde{D}_2(k,n)} \right| = o_P(1).$$

On the set S_n the proof can now be completed as the proof of Theorem 3.2. □

Proof of Theorem 3.3. The decomposition (5.8) yields for $k < k_j$

$$\begin{aligned} V_{k,n}^{(j)}(G) = &-D_1(k,n)D_2(k,n) - A_1(k,n)D_2(k,n) \\ &- A_2(k,n)D_1(k,n) - A_1(k,n)A_2(k,n). \end{aligned}$$

By Assumption A.1(a) an application of Lemma 5.2 conditionally on d_j in addition to an application of the dominated convergence theorem to get the unconditional statement yields on the set M_n as in (5.7) that

$$\max_{1 \le k_j - k \le c\tau^2 d_j^{-2}} \big| A_1(k,n) \big| = o_P(1),$$

$$\max_{1 \le k_j - k \le c\tau^2 d_j^{-2}} \big| A_2(k,n) \big| = O_P(1).$$

Since $\max_{1\le k_j-k\le c\tau^2 d_j^{-2}} |D_1(k,n)| = o_P(1)$ we get on M_n uniformly in $1\le k_j-k\le c\tau^2 d_j^{-2}$

$$\begin{aligned}V_{k,n}^{(j)}(G) &= -\frac{1}{2G}|k_j-k|\big(2G-|k_j-k|\big)d_j^2\\ &\quad -\big(T_{k_j,n}(G;\varepsilon_1,\dots,\varepsilon_n)-T_{k,n}(G;\varepsilon_1,\dots,\varepsilon_n)\big)\big(2G-|k_j-k|\big)\frac{1}{\sqrt{2G}}d_j+o_P(1)\\ &= -|k_j-k|d_j^2-d_j\left(\sum_{i=k+G+1}^{k_j+G}\varepsilon_i+\sum_{i=k-G+1}^{k_j-G}\varepsilon_i-2\sum_{i=k+1}^{k_j}\varepsilon_i\right)+o_P(1).\end{aligned}$$

By stationarity and Assumption A.1(a) we get

$$\begin{aligned}&\left\{d_j\left(\sum_{i=k+G+1}^{k_j+G}\varepsilon_i+\sum_{i=k-G+1}^{k_j-G}\varepsilon_i-2\sum_{i=k+1}^{k_j}\varepsilon_i\right):k=k_j-1,\dots,k_j-c\tau^2d_j^{-2}\right\}\\ &\quad\overset{\mathcal{D}}{=}\left\{U_n(l)=d_j\left(\sum_{i=-l+G+1}^{G}\varepsilon_i+\sum_{i=-l-G+1}^{-G}\varepsilon_i-2\sum_{i=-l+1}^{0}\varepsilon_i\right):1\le l\le c\tau^2d_j^{-2}\right\}.\end{aligned}$$

Note that by the assumption on d_j and Assumption A.5 the three summands are asymptotically independent. Hence by $d_j\xrightarrow{P}0$ and Assumption A.1(a) the functional central limit theorem implies

$$\left\{\frac{U_n(\lfloor s\tau^2d_j^{-2}\rfloor)}{\tau^2}:0\le s\le c\right\}\xrightarrow{D[0,1]}\left\{\sqrt{6}W(s),0\le s\le c\right\},$$

because $\{W_1(s)+W_2(s)-2W_3(s)\}\overset{\mathcal{D}}{=}\{-\sqrt{6}W(s)\}$ for independent standard Wiener processes $\{W_j(\cdot)\}$, $j=1,2,3$ and another standard Wiener process $\{W(\cdot)\}$. More precisely, we first apply the functional central limit theorem given d_j and then get the unconditional assertion above by an application of the dominated convergence theorem. Similar arguments hold for $k_j\ge k$, which implies by $P(M_n)\to 1$ for $-c\le x\le c$

$$\begin{aligned}&P\left(-c\le d_j^2\frac{\hat{k}_j 1_{\{j\le\hat{q}_n\}}-k_j}{\tau^2}\le x\right)\\ &\quad\to P\Big(\max_{-c\le s\le x}\big(-|s|-\sqrt{6}W(s)\big)\ge\max_{x<s\le c}\big(-|s|-\sqrt{6}W(s)\big)\Big)\\ &\quad= P\Big(-c\le\arg\max_{-c\le s\le c}\big(W(s)-|s|/\sqrt{6}\big)\le x\Big).\end{aligned}$$

By Theorem 3.2 and Remark 3.1(a) we get

$$P\left(d_j^2\frac{|\hat{k}_j-k_j|}{\tau^2}\le c\right)\le\big(c^{-\gamma/2}+G^{-\gamma/2}\big)O(1)+o(1)$$

uniformly in n, which becomes arbitrarily small for c large enough. Hence, letting $c \to \infty$ gives assertion (a). Assertion (b) follows because on

$$\{\hat{q}_n = q\} \cap \bigcap_{j=1}^{q_n} \left\{ d_j^2 \frac{|\hat{k}_j - k_j|}{\tau^2} \le c \right\}$$

the change point estimators $\hat{k}_j$, $j = 1, \dots, n$, are asymptotically independent (conditionally on k_j, d_j, $j = 1, \dots, q_n$) by Assumption A.5. Because the above convergence also holds conditionally, this implies (b) by using the dominated convergence theorem in the very last step. □

Proof of Theorem 3.4. Without loss of generality, we can assume that $\sigma = 1$. Let $t_{k,n}(G) = \frac{1}{\sqrt{2G}} \sum_{j=1}^{q_n} (G - |k_j - k|) d_j 1_{\{|k_j - k| < G\}}$ be the value of the statistic if only the signal $\mu_1 + \sum_{j=1}^{q_n} d_j 1_{\{i > k_j\}}$ (without noise) is input into the statistic $T_{k,n}(G)$. Analogously to Lemma A.1 in [19] it holds

$$\begin{aligned} &P\Big(\max_{G \le k \le n-G} \big| T_{k,n}(G) - t_{k,n}(G) \big| \ge \sqrt{4 \log n} \Big) \\ &\le \sum_{k=G}^{n-G} P\big(|Z| \ge \sqrt{4 \log n} \big) \le n \frac{\varphi_Z(\sqrt{4 \log n})}{\sqrt{4 \log n}} \le \frac{C}{2n}, \end{aligned}$$

where $Z \sim N(0, 1)$ and φ_Z is its density. From this it follows that

$$P\Big(\max_{|k-k_j| \ge G, j=1,\dots,q_n} \big| T_{k,n}(G) \big| < c_n \Big) \ge 1 - \frac{C}{2n}$$

as $\max_{|k-k_j| \ge G, j=1,\dots,q_n} |t_{k,n}(G)| = 0$ because the distance between two change points is always greater than $2G$ and $c_n > \sqrt{4 \log n}$. Furthermore,

$$\begin{aligned} &P\Big(\min_{0 \le |k-k_j| < (1-\eta)G, j=1,\dots,q_n} \big| T_{k,n}(G) \big| \ge c_n \Big) \\ &\ge P\Big(\min_{0 \le |k-k_j| < (1-\eta)G, j=1,\dots,q_n} \big| t_{k,n}(G) \big| - \max_{0 \le |k-k_j| < (1-\eta)G, j=1,\dots,q_n} \big| T_{k,n}(G) - t_{k,n}(G) \big| \ge c_n \Big) \\ &\ge 1 - P\Big(\max_{0 \le |k-k_j| < (1-\eta)G, j=1,\dots,q_n} \big| T_{k,n}(G) - t_{k,n}(G) \big| \ge \sqrt{4 \log n} \Big) \ge 1 - \frac{C}{2n} \end{aligned}$$

as $\min_{0 \le |k-k_j| < (1-\eta)G, j=1,\dots,q_n} |t_{k,n}(G)| \ge \eta \sqrt{G} \delta_n / \sqrt{2} \ge c_n + \sqrt{4 \log n}$.

The proof of (a) can now be concluded as in the proof of Theorem 3.1, while (b) can be concluded as in the proof of Theorem 3.2. □

Acknowledgements

This work was supported by DFG grant KI 1443/2-1/2. Most of this work was done while the first author was still at Karlsruhe Institute of Technology (KIT), where her position was financed by the Stifterverband für die Deutsche Wissenschaft by funds of the Claussen-Simon-trust. The authors would like to thank Alexander Meier for pointing out a mistake in a previous version of Theorem 3.2, which lead to Corollary 3.1, as well as for writing the R-package "Mosum", which has been used in the simulations of the revision. Furthermore, they would like to thank four anonymous referees for suggesting to consider the non-asymptotic scenario in Section 3.4 as well as for comments that led to a huge improvement in the simulation section.

Supplementary Material

Additional simulation results (DOI: 10.3150/16-BEJ887SUPP; .pdf). In this supplement, we give some additional simulations illustrating the performance of the above change point estimators in small samples with an emphasis on how the variance estimator and the bandwidth influence the performance.

References

[1] Antoch, J. and Hušková, M. (1999). Estimators of changes. In *Asymptotics*, *Nonparametrics*, *and Time Series*. *Statist*. *Textbooks Monogr.* **158** 533–577. New York: Dekker. MR1724708

[2] Antoch, J., Hušková, M. and Jarušková, D. (2000). Change Point Detection. 5th ERS IASC Summer School.

[3] Bai, J. (1997). Estimating multiple breaks one at a time. *Econometric Theory* **13** 315–352. MR1455175

[4] Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica* **66** 47–78. MR1616121

[5] Bauer, P. and Hackl, P. (1980). An extension of the MOSUM technique for quality control. *Technometrics* **22** 1–7.

[6] Beveridge, S. and Nelson, C.R. (1981). A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the business cycle. *J*. *Monet*. *Econ*. **7** 151–174.

[7] Braun, J.V., Braun, R.K. and Müller, H.-G. (2000). Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika* **87** 301–314. MR1782480

[8] Cappé, O., Moulines, E. and Rydén, T. (2005). *Inference in Hidden Markov Models*. *Springer Series in Statistics*. New York: Springer. MR2159833

[9] Chen, J., Gupta, A.K. and Pan, J. (2006). Information criterion and change point problem for regular models. *Sankhyā* **68** 252–282. MR2303084

[10] Chu, C.-S.J., Hornik, K. and Kuan, C.-M. (1995). MOSUM tests for parameter constancy. *Biometrika* **82** 603–617. MR1366285

[11] Csörgő, M. and Horváth, L. (1997). *Limit Theorems in Change-point Analysis*. Chichester: Wiley. MR2743035

[12] Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *Ann*. *Statist*. **25** 1–37. MR1429916

[13] Davis, R.A., Lee, T.C.M. and Rodriguez-Yam, G.A. (2006). Structural break estimation for nonstationary time series models. *J*. *Amer. Statist*. *Assoc*. **101** 223–239. MR2268041

[14] Dümbgen, L. (1991). The asymptotic behavior of some nonparametric change-point estimators. *Ann*. *Statist*. **19** 1471–1495. MR1126333

[15] Dümbgen, L. and Spokoiny, V.G. (2001). Multiscale testing of qualitative hypotheses. *Ann. Statist.* **29** 124–152. MR1833961

[16] Eichinger, B. and Kirch, C. (2016). Supplement to "A MOSUM procedure for the estimation of multiple random change points." DOI:10.3150/16-BEJ887SUPP.

[17] Franke, J. (2011). Markov switching time series models. In *Handbook of Statistics. Time Series – Methods and Applications* **30** (C.R. Rao and T. Subba Rao, eds.). Amsterdam: Elsevier.

[18] Frick, K., Munk, A. and Sieling, H. (2014). Multiscale change point inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 495–580. MR3210728

[19] Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *Ann. Statist.* **42** 2243–2281. MR3269979

[20] Harchaoui, Z. and Lévy-Leduc, C. (2010). Multiple change-point estimation with a total variation penalty. *J. Amer. Statist. Assoc.* **105** 1480–1493. MR2796565

[21] Hušková, M. (1990). Asymptotics for robust MOSUM. *Comment. Math. Univ. Carolin.* **31** 345–356. MR1077905

[22] Hušková, M. and Slabý, A. (2001). Permutation tests for multiple changes. *Kybernetika* (*Prague*) **37** 605–622. MR1877077

[23] Killick, R., Fearnhead, P. and Eckley, I.A. (2012). Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.* **107** 1590–1598. MR3036418

[24] Kirch, C. (2006). Resampling methods for the change analysis of dependent data. Ph.D. Thesis, University of Cologne.

[25] Kirch, C. and Tajduidje Kamgaing, J. (2016). Detection of change points in discrete valued time series. In *Handbook of Discrete Valued Time Series* (R.A. Davis, S.A. Holan, R.B. Lund and N. Ravishanker, eds.). CRC Press, Boca Raton.

[26] Komlós, J., Major, P. and Tusnády, G. (1975). An approximation of partial sums of independent RV's and the sample DF. I. *Probab. Theory Related Fields* **32** 111–131.

[27] Komlós, J., Major, P. and Tusnády, G. (1976). An approximation of partial sums of independent RV's, and the sample DF. II. *Probab. Theory Related Fields* **34** 33–58. MR0402883

[28] Kuelbs, J. and Philipp, W. (1980). Almost sure invariance principles for partial sums of mixing B-valued random variables. *Ann. Probab.* **8** 1003–1036. MR0602377

[29] Lange, T. and Rahbek, A. (2009). An introduction to regime switching time series models. In *Handbook of Financial Time Series* (T.G. Andersen, R.A. Davis, J.-P. Kreiß and T. Mikosch, eds.) Springer, Berlin.

[30] Ling, S. (2007). Testing for change points in time series models and limiting theorems for NED sequences. *Ann. Statist.* **35** 1213–1237. MR2341704

[31] Liu, J., Wu, S. and Zidek, J.V. (1997). On segmented multivariate regression. *Statist. Sinica* **7** 497–525. MR1466692

[32] Marušiaková, M. (2009). Tests for multiple changes in linear regression models. Ph.D. Thesis, Charles University, Prague.

[33] Messer, M., Kirchner, M., Schiemann, J., Roeper, J., Neininger, R. and Schneider, G. (2014). A multiple filter test for the detection of rate changes in renewal processes with varying variance. *Ann. Appl. Stat.* **8** 2027–2067. MR3292488

[34] Muggeo, V.M. and Adelfio, G. (2010). Efficient change point detection for genomic sequences of continuous measurements. *Bioinformatics* btq647.

[35] Muhsal, B. (2013). Change-point methods for multivariate autoregressive models and multiple structural breaks in the mean. Ph.D. Thesis, Karlsruhe Institute of Technology, Karlsruhe.

[36] Pan, J. and Chen, J. (2006). Application of modified information criterion to multiple change point problems. *J. Multivariate Anal.* **97** 2221–2241. MR2301636

[37] Phillips, P.C.B. and Solo, V. (1992). Asymptotics for linear processes. *Ann. Statist.* **20** 971–1001. MR1165602

[38] Politis, D.N. and Romano, J.P. (1995). Bias-corrected nonparametric spectral estimation. *J. Time Series Anal.* **16** 67–103. MR1323618

[39] Preuss, P., Puchstein, R. and Dette, H. (2015). Detection of multiple structural breaks in multivariate time series. *J. Amer. Statist. Assoc.* **110** 654–668. MR3367255

[40] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

[41] Stout, W.F. (1974). *Almost Sure Convergence*. New York: Academic Press. MR0455094

[42] Tong, H. (1990). *Non-linear Time Series*: *A Dynamical System Approach*. London: Oxford Univ. Press.

[43] Vostrikova, L.J. (1981). Detecting disorder in multidimensional random processes. *Sov. Math. Dokl.* **24** 55–59.

[44] Yao, Y.-C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statist. Probab. Lett.* **6** 181–189. MR0919373

[45] Yao, Y.-C. and Au, S.T. (1989). Least-squares estimation of a step function. *Sankhyā Ser. A* **51** 370–381. MR1175613

[46] Yau, C.Y. and Zhao, Z. (2016). Inference for multiple change-points in time series via likelihood ratio scan statistics. *J. Roy. Statist. Soc. Ser. A* **78** 895–916.

[47] Yokoyama, R. (1980). Moment bounds for stationary mixing sequences. *Probab. Theory Related Fields* **52** 45–57. MR0568258

Received August 2015 and revised July 2016