A MOSUM procedure for the estimation of multiple random change points

Claudia Kirch† Birte Muhsal‡

January 9, 2014

Abstract

In this work, we investigate statistical properties of change point estimators based on moving sum statistics, where we allow for random exogenous change points as are e.g. considered in Hidden Markov or regime switching models extending results for testing in a classical (deterministic) multiple change point situations. 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 cannot be improved in general by deriving the limit distribution of properly scaled estimators under somewhat stronger assumptions. 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.

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

AMS Subject Classification 2010: 62G05; 62G20; 62M10

1 Introduction

There are essentially two approaches to multiple change point problems: Model selection and hypothesis testing. Most papers concerned with hypothesis testing in the context of change point problems propose tests for a fixed or bounded number of changes

---

*This work was supported by DFG grant KI 1443/2-1/2 and the Stifterverband für die Deutsche Wissenschaft by funds of the Claussen-Simon-trust who also financed the position of the first author.

†Karlsruhe Institute of Technology (KIT), Institute of Stochastics, Kaiserstr. 89, 76133 Karlsruhe, Germany; claudia.kirch@kit.edu

‡Karlsruhe Institute of Technology (KIT), Institute of Stochastics, Kaiserstr. 89, 76133 Karlsruhe, Germany; birte.muhsal@kit.edu
1 Introduction

(confer [Bai and Perron, 1998] for such $F$-type tests or [Marušiaková, 2009] 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 [Csörgő and Horváth, 1997]). 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 [Vostrikova, 1981], 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. An advantage of the binary segmentation procedure is its simplicity and computational efficiency, one drawback the inability to control the overall significance level. 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 [Fryzlewicz, 2013]. Another recent fully parametric approach [Frick et al., 2014] minimizes the number of change points over the acceptance region of a suitable multiscale test.

Model selection was first proposed by [Yao, 1988] who used Schwarz’s criterion ([Schwarz, 1978]) 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 ([Yao and Au, 1989] in context of mean changes or [Liu et al., 1997] for multivariate regression models), least absolute deviations ([Braun et al., 2000] for mean and simultaneous variance changes or [Bai, 1997] for linear regression models) or the minimum description length (confer [Davis et al., 2006] 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 e.g. [Killick et al., 2012], [Chen et al., 2006] and [Pan and Chen, 2006] 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 unnecessary complexity of the model, when structural breaks appear close to each other, at the beginning or at the end of the data set.

In this paper, we investigate properties of change point estimators based on moving sum statistics, first proposed but not mathematically analyzed by [Antoch et al., 2000]. This method does not require to fix an upper bound for the number of changes, is not computationally expensive and the overall significance level is controlled. On the downside, the performance depends crucially on a bandwidth choice, which ideally should be chosen as the minimum distance between two neighboring change points. This method is useful as a diagnostic tool as the statistics and corresponding change point estimators can easily be calculated for several bandwidths due to computational simplicity, where the corresponding plots give valuable insights for the change analysis. In the context of testing, such MOSUM statistics have already been investigated by [Bauer and Hack, 1980], [Chu et al., 1995] or [Hušková, 1990] as well as [Hušková and Slabý, 2001]. More recently, [Preuß et al., 2013] use moving sum statistics in the frequency domain to detect changes in the autocovariance structure of multivariate time series. The focus in this paper is not on testing but on the use of such moving sum statistics to estimate the number and locations of the change points. We will not only show consistency of the corresponding change point estimators for the classical multiple change situation but also consider the asymptotic properties of these estimators for random change points including regime switching models (for a recent survey we refer to [Franke, 2011]. Regime-switching models are a very flexible method to model structural breaks in time series still allowing for meaningful statistical inference. 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 like from a classical multiple change point model with deterministic (but unknown) change points. Therefore, it is not surprising that
classical change point methods can help to analyze the properties of regime switching models. We will demonstrate this fact by using a simple mean change situation, where only the conditional mean is governed by an unobservable process independent of the stationary error sequence. While this is the easiest framework for changes, it still shows the technical difficulties and solutions that arise. Furthermore, many change point statistics even for more complex statistics are based on partial sums, similarly to score type statistics or more generally statistics based on estimating functions (confer Kirch and Tajdude-Kamgaing, 2014 for details).

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 a first subsection we introduce the multiple change model that we will be using throughout the paper including possibly random change points 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 can in general not be improved in Section 3.3. 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 respectively 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

\[ X_i = \sum_{j=1}^{q+1} \mu_j 1_{\{k_{j-1,n} < i \leq k_{j,n}\}} + \varepsilon_i, \quad i = 1, ..., n, \]

where

\[ 0 = k_{0,n} < k_{1,n} = [\vartheta_1 n] \leq ... \leq k_{q,n} = [\vartheta_q n] \leq k_{q+1,n} = n, \quad 0 < \vartheta_1 \leq ... \vartheta_q \leq 1. \]

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

In this model, the distance between change points grows with rate \( n \). While the increasing distance is a necessary condition to show asymptotic power one of the test respectively consistency of the estimators, the rate can be relaxed (confer Theorems 2.2 and 3.1).
2 MOSUM tests for multiple changes

A regime-switching model, which describes changes in the mean, is given by

\[ X_i = \tilde{\mu}_{Q_i} + \tilde{\epsilon}_i, \quad i = 1, \ldots, n, \]

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 \( \tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_n \) 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 [Lange and Rahbek, 2009]) such as threshold models introduced by [Tong, 1990] 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 [Dahlhaus, 1997] we adopt a different approach for asymptotics, assuming an underlying triangular scheme

\[ X_i^{(n)} = \tilde{\mu}_{Q_i^{(n)}} + \tilde{\epsilon}_i^{(n)} \]

with a sequence of non-observable processes \( \{Q_i^{(n)} : 1 \leq i \leq n\} \). Furthermore, we allow for both a fixed and increasing number of change points \( q_n \), where the distance between adjacent change points grows at least with the same rate as the bandwidth of the moving sum statistic (see (2.6) below). In small samples, this translates to a low tendency to switch and only relatively few change points.

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 \( k_{q_n+1,n} = n \))

\[ X_i = \sum_{j=1}^{q_n+1} \mu_{j,n} 1\{k_{j-1,n} < i \leq k_{j,n}\} + \tilde{\epsilon}_i, \quad \mu_{j,n} \in \{\tilde{\mu}_{1,n}, \ldots, \tilde{\mu}_{K_n,n}\}, \]

\( i = 1, \ldots, n \), where now 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

\[ X_i = \mu_{1,n} + \sum_{j=1}^{q_n} d_{j,n} 1\{i > k_{j,n}\} + \tilde{\epsilon}_i, \quad d_{j,n} = \mu_{j+1,n} - \mu_{j,n}, \quad (2.1) \]

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 \( \mu_j \) for notational ease.

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

**Assumption E.1.** The number of changes, changes and their locations \( \{q_n, d_{j,n}, k_{i,n}, i, j = 1, \ldots, q_n\} \) are independent of the error sequence \( \{\tilde{\epsilon}_i\} \).

This assumption 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. For the classical change point setting this is no additional condition.
Assumption E.2. There exists a standard Wiener process \( \{W(k) : 1 \leq k \leq n\} \) and \( \nu > 0 \) such that (possibly after changing the probability space)
\[
\sum_{i=1}^{n} \varepsilon_i - \tau W(n) = O(a_1^{1/(2+\nu)}) \quad \text{a.s.}
\]
with an existing and strictly positive long-run variance
\[
\tau^2 = \sigma^2 + 2 \sum_{h>0} \gamma(h) > 0, \quad \gamma(h) = \text{cov}(\varepsilon_0, \varepsilon_h).
\]

Such invariance principles have first been derived for i.i.d. random variables by Komlós et al., 1975 and Komlós et al., 1976. Subsequently, many classes of time series have also been considered, e.g. mixing time series (Theorem 4 in Kuelbs and Philipp, 1980) or near-epoch dependent time series (Ling, 2007).

2.2 Testing for multiple mean changes using a moving sum statistic

Hušková and Slabý, 2001 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 null distribution plays an important role in this analysis. Consequently, we repeat the relevant results extending them to depend errors as in E.2 and show that this test statistic has asymptotic power one under the more general regime switching model above. Consider the following moving sum statistic

\[
T_n(G) = \max_{G \leq k \leq n-G} \frac{|T_{k,n}(G)|}{\tau},
\]

\[
T_{k,n}(G) = T_{k,n}(G; X_1, \ldots, X_n) = \frac{1}{\sqrt{2G}} \left( \sum_{i=k+1}^{k+G} X_i - \sum_{i=k-G+1}^{k} X_i \right),
\]

with bandwidth \( G = G(n) \) fulfilling
\[
\frac{n}{G} \longrightarrow \infty \quad \text{and} \quad \frac{n^{3/2} \log n}{G} \longrightarrow 0,
\]

where \( \nu \) and \( \tau \) are as in E.2. The bandwidth assumption guarantees that \( G \) converges to infinity but not too fast.

The statistic in (2.2) compares at every time point \( G \leq k \leq n - G \) the mean of the subsample \( X_{k-G}, \ldots, X_k \) with the mean of the subsample \( X_{k+1}, \ldots, X_{k+G} \), where a large difference indicates a change at this point. The following theorem gives the null asymptotics of the test statistic quantifying the acceptable deviation of the difference from zero.

Theorem 2.1. Let the null hypothesis hold, i.e. \( X_i = \mu_1 + \varepsilon_i, i = 1, \ldots, n \), with \( \{\varepsilon_i\} \) fulfilling E.2. If the bandwidth \( G \) fulfills (2.3), then
\[
a(n/G) T_n(G) - b(n/G) \overset{D}{\longrightarrow} \Gamma,
\]

where \( \Gamma \) follows a Gumbel extreme value distribution, i.e. \( P(\Gamma \leq x) = \exp(-2 \exp(-x)) \) and
\[
a(x) = \sqrt{2 \log x}, \quad b(x) = 2 \log(x) + \frac{1}{2} \log \log x + \log(3/2) - \frac{1}{2} \log \pi.
\]

The assertion remains true if \( \tau \) 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
\[
\max_{G \leq k \leq n-G} |\hat{\tau}_{k,n}^2 - \tau^2| = o_P \left( (\log(n/G))^{-1} \right).
\]
Consequently, we get an asymptotic level \( \alpha \) test if the null hypothesis is rejected for \( T_n(G) > D_n(G; \alpha) \) with

\[
D_n(G; \alpha) = \frac{b(n/G) + c_\alpha}{a(n/G)}, \quad c_\alpha = -\log \log \frac{1}{\sqrt{1-\alpha}}. \tag{2.5}
\]

We will now prove, that this test statistic rejects with asymptotic power one under the alternative with possibly random changes as in (2.1) which includes the classic change point alternatives.

**Theorem 2.2.** Let \( X_1, \ldots, X_n \) follow (2.1) with

\[
P\left( \min_{0 \leq j \leq q_n} |k_{j+1} - k_j| > 2G \right) \to 1, \tag{2.6}
\]

\[
P\left( \min_{G \leq k \leq n-G} \hat{\tau}_{k,n} > 0 \right) \to 1, \tag{2.7}
\]

and

\[
\frac{1}{\min_{1 \leq j \leq q_n} d_j^2} = o_P\left( \frac{G}{\log(n/G)} \right). \tag{2.8}
\]

Furthermore, let \( \{\varepsilon_t\} \) fulfill the assumptions of Theorem 2.1 and \( E[T] \) hold. Then, we get for any \( c \in \mathbb{R} \)

\[
P(a(n/G)T_n(G) - b(n/G) \geq c) \to 1,
\]

i.e. the test has asymptotic power one. The assertion remains true, if \( \tau \) is replaced by an estimator \( \hat{\tau}_{k,n} \) which can depend on \( k \) such that

\[
\max_{G \leq k \leq n-G} \hat{\tau}_{k,n}^2 = o_P\left( \frac{G \min_{1 \leq j \leq q_n} d_j^2}{\log(n/G)} \right). \tag{2.9}
\]

Assumption (2.7) may seem unnecessary at first, but while the Bartlett weights below guarantee that the estimator is non-negative this is not necessarily the case for the flat-top kernel below. As this causes a problem in many statistical procedures the standard solution is to take the maximum of the estimator and the sample variance divided by the logarithm of the sample size (the latter to guarantee scale invariance in addition to asymptotic consistency).

Assumption (2.8) allows in particular for local changes. In the next section we propose estimators for the long-run variance which fulfill the above assumptions for an appropriate bandwidth choice.

### 2.3 Estimating the asymptotic variance

The small sample power as well as precision of estimators depends crucially on how the (long-run) variance \( \tau^2 \) 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 leading to too large an estimate (confer Figure 2.1). 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 CUSUM statistics estimates a possible
2 MOSUM tests for multiple changes

change (using the point of maximum of the statistic) and then calculates the sample variance after a separate mean correction for both subsamples

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

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

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

where \(\bar{X}_{l,j} = \frac{1}{j-l+1} \sum_{i=l}^{j} X_i\), as variance estimator. Figure 2.1 illustrates the behavior of this estimator in comparison to the sample variance 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 (2.10) (solid line), where the true variance is 1 (as indicated by the dotted horizontal line) for different choices of bandwidths. 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.

Similarly, we propose to use time dependent MOSUM versions of flat-top kernels as proposed by [Politis and Romano, 1995] for the long-run variance in case of dependent errors

\[
\hat{\tau}^2_{k,n} := \gamma_k(0) + 2 \sum_{h=1}^{\Lambda_n} \omega(h/\Lambda_n)\hat{\gamma}_k(h)
\]
2 MOSUM tests for multiple changes

with autocovariance estimator

\[
\hat{\gamma}_k(h) := \frac{1}{2G} \sum_{i=k-G+1}^{k-h} (X_i - \bar{X}_{k-G+1,k})(X_{i+h} - \bar{X}_{k-G+1,k}) + \frac{1}{2G} \sum_{i=k+1}^{k+G-h} (X_i - \bar{X}_{k+1,k+G})(X_{i+h} - \bar{X}_{k+1,k+G}),
\]

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

\[
w(t) = \begin{cases} 
1, & |t| \leq 1/2, \\
2(1 - |t|), & 1/2 < |t| < 1, \\
0, & |t| \geq 1.
\end{cases}
\]

Remark 2.1. If the observations are allowed to change simultaneously in mean and variance, then the above two-sided MOSUM approach to estimation will result in estimators that are larger than the smaller variance. In this case, power and precision of estimators can possibly be further increased by using the minimum of the left and right-sided estimators at the cost of a larger estimation error as less observations are included in the estimation. For i.i.d. data, e.g. such estimators are defined by

\[
\hat{\sigma}^2_{k,n,m} = \min(\hat{\sigma}^2_{k,n,l}, \hat{\sigma}^2_{k,n,r}) = \min \left( \frac{1}{G} \sum_{i=k-G+1}^{k} (X_i - \bar{X}_{k-G+1,k})^2, \frac{1}{G} \sum_{i=k+1}^{k+G} (X_i - \bar{X}_{k+1,k+G})^2 \right).
\]

For this estimator the same rate as in Theorems 2.3 and 2.4 holds by analogous arguments. More details and some empirical results can be found in Section 4.

Theorem 2.3. For \( X_i = \mu + \varepsilon_i, i = 1, \ldots, n \), with \( \{\varepsilon_i\} \) fulfill Assumption E.2 as well as \( \mathbb{E}|\varepsilon_i|^4 < \infty \) and

\[
\sup_{h \in \mathbb{Z}} \sum_{k \in \mathbb{Z}} \sum_{t \in \mathbb{Z}} |\nu(h, k, l)| < \infty, \quad (2.11)
\]

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

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

\[
\max_{G \leq k \leq n-G} |\hat{\sigma}^2_k - \sigma^2| = O_P \left( \frac{n^{1/2}}{G} \right).
\]

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

\[
\max_{G \leq k \leq n-G} |\hat{\tau}^2_k - \tau^2| = O_P \left( \frac{\Lambda_n n^{1/2}}{G} + r_n \right),
\]

where

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

Remark 2.2. a) We get the weaker rate \( O_P(n^{2/(2+\nu)}/G) \) if only \( \mathbb{E}|X_i|^{2+\nu} < \infty, \ 0 < \nu < 2, \) and the stronger rate \( O_P(\sqrt{\log(n/G)}/G) \) for \( \nu > 2 \) (for details we refer to [Muhsal, 2013], 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 $\sum_{h \in \mathbb{Z}} h^\alpha |\gamma(h)| < \infty$, $0 < \alpha \leq 1$, then for both the Bartlett as well as flat-top weights $r_n = \Lambda_n^{-\alpha}$.

If changes are present, the estimators are influenced by these changes as demonstrated in Figure 2.1 but we still obtain:

**Theorem 2.4.** Let the error sequence $\{\varepsilon_i\}$ fulfill the assumptions of Theorem 2.3 in addition to $\max_{2 \leq j \leq q_n + 1} |\mu_j - \mu_1| = O_P(1)$. Then, we get for the random change point model (2.1) (which includes the classical model):

a) If $n/G^2 \to 0$,
$$\max_{G \leq k \leq n-G} \hat{\sigma}^2_{k,n} = O_P(1).$$

b) If $\Lambda_n^2 n/G^2 \to 0$,
$$\max_{G \leq k \leq n-G} \hat{\tau}^2_{k,n} = O_P(\Lambda_n).$$

If (2.6) holds, we can replace the condition $\max_{2 \leq j \leq q_n + 1} |\mu_j - \mu_1| = O_P(1)$ by $\max_{1 \leq j \leq q_n + 1} d_j^2 = O_P(1)$ (for details we refer to the proof of Theorem 6.18 in [Muhsal, 2013]).

### 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 show asymptotic consistency for both estimators. Additionally, we will derive rates of convergence of the estimator for the locations of the changes and show that these rates cannot be improved in general.

#### 3.1 Asymptotic consistency

Figure 3.1 shows a time series with i.i.d. errors and three change points marked by the 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 (2.10). 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.

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

\[
\hat{\tau}_{k,n} - 1 \geq D_n(G; \alpha_n) \quad \text{for} \quad k = v_j, \ldots, w_j, \\
\hat{\tau}_{k,n} - 1 < D_n(G; \alpha_n) \quad \text{for} \quad k = v_j - 1, w_j + 1, \\
w_j - v_j \geq \eta G
\]

with \( 0 < \eta < 1/2 \) arbitrary but fixed, \( (3.1) \)

where \( D_n(G; \alpha) \) is as in \( (2.5) \). Then,

\[
\hat{q}_n = \text{the number of pairs} \ (v_j, w_j) \quad (3.2)
\]

is an estimator of the number of change points \( q \), while

\[
\hat{k}_j := \arg \max_{v_j \leq k \leq w_j} \frac{|T_{k,n}(G)|}{\hat{\tau}_{k,n}} \quad (3.3)
\]

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, i.e. when the statistic crosses the critical line.

This method has already been proposed but not mathematically analyzed by [Antoch et al., 2000].

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 cannot be improved in general.

**Theorem 3.1.** Let the assumptions of Theorem 2.2 hold. Additionally, let

\[
\max_{|k - k_j| \geq G} |\hat{\tau}_{k,n}^2 - \tau^2| = o_P \left( (\log(n/G))^{-1} \right),
\]

\( c_{\alpha_n} \) as in \( (2.5) \) and a sequence \( \alpha_n \) fulfilling

\[
\alpha_n \to 0, \quad \frac{c_{\alpha_n}}{a(n/G)} = O(1),
\]

where \( a(\cdot) \) is as in Theorem 2.1. Then, it holds as \( n \to \infty \)

a) \( P(\hat{q}_n = q_n) \to 1. \)

b) \( P \left( \max_{1 \leq j \leq \hat{q}_n} |\hat{k}_j| 1_{\{j \leq \hat{q}_n\}} - k_j \geq G \right) \to 0. \)

**Remark 3.1.** The first assumption on the (long-run) variance estimator holds e.g. for translation invariant estimators, i.e. \( \hat{\tau}_{k,n}^2 = \hat{\tau}_{k,n}^2(X_{k-G+1}, \ldots, X_{k+G}) = \hat{\tau}_{k,n}^2(X_{k-G+1} - \mu, \ldots, X_{k+G} - \mu) \) for any \( \mu \), if additionally

\[
\max_{G \leq k \leq n-G} |\hat{\tau}_{k,n}^2(\varepsilon_1, \ldots, \varepsilon_n) - \tau^2| = o_P \left( (\log(n/G))^{-1} \right).
\]
3 MOSUM-based estimators for multiple change points

3.2 Convergence rates

In order to get better rates of convergence, we need the following bound for higher moments of sums of the error sequence:

**Assumption E.3.** For some $\gamma > 2$ and some constant $C > 0$ it holds for any $-\infty < \ell \leq u < \infty$

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

This assumption holds e.g. for independent random variables or linear processes, for which it follows from the Beveridge-Nelson decomposition (confer [Beveridge and Nelson, 1981, Phillips and Solo, 1992]). It also holds for martingale difference sequences, certain $\Phi$-mixing sequences (confer [Stout, 1974], Theorem 3.7.8) but also for certain $\alpha$-mixing sequences (confer [Yokoyama, 1980], Theorem 1). We need this assumption to get the following forward as well as backward Hájek-Rényi-type inequalities:

**Lemma 3.1.** Under Assumption E.3 it holds for any positive and non-increasing sequence $\{c_k\}$, any $1 \leq \ell \leq u$, any $m \in \mathbb{Z}$ and any $\delta > 0$

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

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

where $\tilde{C}$ only depends on $C$ and $\gamma$ of Assumption E.3.

Not only does this inequality yield stochastic convergence rates for weighted maxima of partial sums, it also allows for a trade-off between the stochastic convergence rates and the rate of convergence of the probability. This will be essential in the proof of part b) of Theorem 3.2 below. While Assumption E.2 also implies the stochastic convergence rates of the (forward) Hájek-Rényi-inequalities a fact we already made extensive use of in the previous results, it does not allow for this trade-off.

**Theorem 3.2.** Let the assumptions of Theorem 3.1 hold, in addition to E.3 and

$$P \left( \min_{1 \leq j \leq q_n} |d_j| | \geq \delta_n \right) \rightarrow 1.$$

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

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

where the rates do not depend on $j$.

b) If $P(q_n > \gamma_n) \rightarrow 0$, then

$$P \left( \max_{1 \leq j \leq \hat{q}_n} |k_j 1_{(j \leq \hat{q}_n)} - k_j| \geq \xi_n \right) = \gamma_n \delta_n^{-\gamma} \left( \xi_n^{-\hat{\gamma}} + G^{-\hat{\gamma}} \right) O(1) + o(1).$$

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 $j$).
3 MOSUM-based estimators for multiple change points

Remark 3.2. a) If the probability on the left-hand side of a) is replaced by the conditional probability (given \(d_j\)), then \(\delta_n\) on the right-hand side can be replaced by \(d_j\).

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

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

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

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

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

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, because for i.i.d. errors, \(d_j = d\), \(\hat{\tau}_{j,n} = \tau\) and \(q_n\) deterministic change points \((\hat{k}_j 1_{(j < \hat{q}_n)} - k_j)\), \(j = 1, \ldots, q_n\) are independent and identically distributed on the asymptotic 1-set \(M_n\) (as in (5.7)), hence

\[ P\left( \max_{1 \leq j \leq q_n} |\hat{k}_j 1_{(j < \hat{q}_n)} - k_j| > \xi_n \right) = 1 - (1 - P(|\hat{k}_1 - k_1| > \xi_n))^{q_n} \]

\[ = 1 - \exp[q_n \log(1 - P(|\hat{k}_1 - k_1| > \xi_n))] \]

\[ = 1 - \exp[-q_n(P(|\hat{k}_1 - k_1| > \xi_n) + o(P(|\hat{k}_1 - k_1| > \xi_n)))], \]

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.

3.3 Asymptotic distribution

In this section, we derive the asymptotic distribution of the estimators for the locations for local changes. 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 [Antoch and Hušková, 1999] as well as [Dümbgen, 1991]). In particular this shows that the convergence rates of the estimators cannot be improved in general (confer also Remark 3.2).

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

Assumption E.4. 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 (3.3) below,
Simulations

The change point estimators are asymptotically independent. The forward functional central limit theorem is fulfilled by Assumption E which follows for example for exponentially mixing sequences under moment conditions (confer [Kuelbs and Philipp, 1980]). 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 the assumptions of Theorem 3.2 hold but replace (2.6) by the existence of a $c > 2$ such that

$$P \left( \min_{0 \leq k \leq q} |k_{j+1} - k_j| > cG \right) \rightarrow 1$$

(3.5)

a) If $d_j \xrightarrow{p} 0$ but $d_j^2 G \xrightarrow{p} \infty$, then it holds for $j = 1, \ldots, q_n$ as $n \rightarrow \infty$

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

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

b) If additionally Assumption E(i) or (ii) holds, then for a fixed number of changes $q_n = q$ and

$$\max_{j=1, \ldots, q} d_j^2 \rightarrow 0, \quad G \min_{j=1, \ldots, q} d_j^2 \rightarrow \infty,$$

it holds as $n \rightarrow \infty$

$$\tau^{-2} \left( d_1^2 (\hat{k}_1 1_{(1 \leq q_n)} - k_1), \ldots, d_q^2 (\hat{k}_q 1_{(q \leq q_n)} - k_q) \right) \xrightarrow{D} (S_1, \ldots, S_q),$$

where

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

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

Simulations

In this section we illustrate 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. Nonparametric estimators for the long-run variance are relatively imprecise even for larger sample sizes as this is a difficult statistical estimation problem. Consequently, we use i.i.d. errors to illustrate the influence of the important novelties in this paper which are a) using moving sums for estimation and b) using moving sum variance estimators. 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 this influences the estimators.

We consider the situation of independent and identically standard normally distributed random errors and simulate a random sample $X_1, \ldots, X_n$ of size $n = 500$. The test statistic $T_{k,n}(G)/\hat{\sigma}_{k,n}$ is calculated for four different bandwidths $G = 25, 40, 50, 75$ and $\hat{\sigma}_{k,n} = \sigma$ the true value, $\hat{\sigma}_{k,n} = \hat{\sigma}$ the sample variance as well as $\hat{\sigma}_{k,n}$ as in (2.10). Further, we set $\eta = 0.15$ (refer to (3.1)), choose the level $\alpha = 0.05$ and compute the critical value $D_n(G; \alpha)$ as in (2.5). The results can be found in Figures 4.1, 4.3. It becomes apparent that using the standard sample variance as in the second columns results in a poor performance of the estimation due to the contamination of the variance estimators by the mean changes. In contrast the moving sum variance estimator as
4 Simulations

Figure 4.1: Sample path and performance of $\sigma^{-1}T_{k,n}(G)$ (first column), $\hat{\sigma}^{-1}T_{k,n}(G)$ (second column) and $\hat{\sigma}_{h,1}^{-1}T_{k,n}(G)$ (third column) with bandwidths $G = 25, 40, 50, 75$ for independent normal distributed errors with variance 1.

Figure 4.2: Sample path and performance of $\sigma^{-1}T_{k,n}(G)$ (first column), $\hat{\sigma}^{-1}T_{k,n}(G)$ (second column) and $\hat{\sigma}_{h,1}^{-1}T_{k,n}(G)$ (third column) with bandwidths $G = 25, 40, 50, 75$ for independent normal distributed errors with variance 1.

Figure 4.3: Sample path and performance of $\sigma^{-1}T_{k,n}(G)$ (first column), $\hat{\sigma}^{-1}T_{k,n}(G)$ (second column) and $\hat{\sigma}_{h,1}^{-1}T_{k,n}(G)$ (third column) with bandwidths $G = 25, 40, 50, 75$ for independent $t$-distributed errors with 3 degrees of freedom.
4 Simulations

in the third column performs quite well, it even behaves somewhat better than using the true variance as in the first column because the peaks are more pronounced due to the overestimation of the variance close to the true change point but not right at the change point.

Obviously, bigger changes are easier detected than smaller changes and large bandwidths also result in better estimates as long as the changes are further apart than twice the bandwidth guaranteeing that only one change point is included in the calculation of $T_{k,n}$ for every $k$. This is illustrated in Figure 4.2 where the changes are closer. Here, the smaller bandwidths $G = 25, 50$ have still three clear peaks, but with growing bandwidth the first two peaks merge so that only one of them is detected by the procedure. This is very similar to using CUSUM type statistics in a multiple change situation.

In practise a mean change is often accompanied by a change in variance. In this case, adapted moving variance estimators as in Remark 2.1 are quite useful as demonstrated by Figure 4.4. As soon as the variance difference is too big (first two columns) the moving sum estimator does no longer work well as it estimates a convex combination of both variances. The moving estimation based only on the left (second row) respectively right (third row) window works well if the variance increases respectively decreases but not otherwise. However, using the minimum of the two as introduced in Remark 2.1 yields satisfactory results in all three cases. This finding is also confirmed by the simulations in Figure 4.5 which gives a more realistic scenario.

To conclude the MOSUM procedure can be used to get consistent estimators for the number and location of change points but may be even more useful as a diagnostic tool if applied with different bandwidths as demonstrated in the above plots.
4 Simulations

Figure 4.4: Sample paths and performances of $\hat{\sigma}^{-1}_{k,n}T_{k,n}(G)$ (second row), $\hat{\sigma}^{-1}_{k,n,l}T_{k,n}(G)$ (third row), $\hat{\sigma}^{-1}_{k,n,r}T_{k,n}(G)$ (fourth row) as well as $\hat{\sigma}^{-1}_{k,n,m}T_{k,n}(G)$ (fifth row).

Figure 4.5: Sample paths and performances of $\hat{\sigma}^{-1}_{k,n}T_{k,n}(G)$ (second row), $\hat{\sigma}^{-1}_{k,n,l}T_{k,n}(G)$ (third row), $\hat{\sigma}^{-1}_{k,n,r}T_{k,n}(G)$ (fourth row) as well as $\hat{\sigma}^{-1}_{k,n,m}T_{k,n}(G)$ (fifth row).
5 Proofs

In this section we prove the results of the previous sections.

5.1 Proofs of Section 2

Proof of Theorem 2.1. By the invariance principle in $E_2$ and (2.3) we get
\[
\max_{G \leq k \leq n-G} \frac{1}{\sqrt{2G}} \sum_{i=k+1}^{k+G} \varepsilon_i - (W(k + G) - W(k)) \mid = O_P \left( \frac{n^{1/(2+\nu)}}{2G} \right) = o_P \left( a(n/G)^{-1} \right).
\]

Consequently, we can replace $X_i/\tau$ 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 [Hušková and Slabý, 2001]. This implies
\[
\max_{G \leq k \leq n-G} \mid T_{k,n}(G) \mid = O_P (\sqrt{\log(n/G)})
\]
showing that we can replace $\tau$ by $\hat{\tau}_{k,n}$ without changing the asymptotics since by assumption (2.4) it holds
\[
\max_{G \leq k \leq n-G} \left\| 1 - \hat{\tau}_{k,n} \right\| = o_P \left( (\log(n/G))^{-1} \right).
\]

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
\[
\max_{G \leq k \leq n-G} \left\| \tilde{\varepsilon}_{k,n}^2 \right\| = o_P \left( \min_{j=1,...,q_n} d_j^2 \frac{\log(n/G)}{c_{\alpha_n}} \right),
\]
then
\[
P \left( \min_{0 \leq |k-j| < (1-\varepsilon)G \atop j=1,...,q_n} \frac{|T_{k,n}(G)|}{\hat{\tau}_{k,n}} < D_n(G;\alpha_n) \right) \rightarrow 0,
\]
where $c_{\alpha_n}, D_n(G;\alpha_n)$ as in (2.5).

b) If (as in Theorem 3.1)
\[
\max_{|k-j| \geq G, j=1,...,q_n} |\varepsilon_{k,n}^2 - \tau^2| = o_P \left( (\log(n/G))^{-1} \right),
\]
then it holds uniformly in $\alpha$
\[
P \left( \max_{|k-j| \geq G, j=1,...,q_n} \frac{|T_{k,n}(G)|}{\hat{\tau}_{k,n}} \geq D_n(G;\alpha) \right) \leq \alpha.
\]
5 Proofs

Proof. To obtain a), note that by assumption it is sufficient to prove the assertion if additionally \( \min_{0 \leq j \leq q_n} |k_{j+1} - k_j| > 2G \) holds. In this case some calculations (for details we refer to [Muhsal, 2013], proof of Theorem 6.1) give for \( 0 \leq |k - k_j| \leq (1-\varepsilon)G \)
\[
T_{k,n}(G; X_1, \ldots, X_n) = T_{k,n}(G; \varepsilon_1, \ldots, \varepsilon_n) + \frac{d_j}{\sqrt{2G}} (G - |k - k_j|).
\]

Hence, by (5.1)
\[
\min_{0 \leq |k - k_j| < (1-\varepsilon)G} |T_{k,n}(G)| = \min_{0 \leq |k - k_j| < (1-\varepsilon)G} |T_{k,n}(G; \varepsilon_1, \ldots, \varepsilon_n) + \frac{d_j}{\sqrt{2G}} (G - |k - k_j|) |
\geq \min_{1 \leq j \leq q_n} |d_j| \sqrt{\frac{\varepsilon}{2}} - \max_{G \leq k \leq n - G} |T_{k,n}(G; \varepsilon_1, \ldots, \varepsilon_n)|
= \min_{1 \leq j \leq q_n} |d_j| \sqrt{\frac{\varepsilon}{2}} + O_P(\sqrt{\log(n/G)}).
\]

We conclude
\[
P \left( \min_{0 \leq |k - k_j| < (1-\varepsilon)G} \frac{|T_{k,n}(G)|}{\tilde{\tau}_{k,n}} < D_n(G, \alpha_n), \min_{0 \leq j \leq q_n} |k_{j+1} - k_j| > 2G, \min_{G \leq k \leq n - G} \tilde{\tau}_{k,n} > 0 \right)
\leq P \left( \min_{1 \leq j \leq q_n} |d_j|^{-1} G^{-1/2} \left( \max_{G \leq k \leq n - G} \frac{\varepsilon_{n, \alpha_n} + b(n/G)}{a(n/G)} + O_P(\sqrt{\log(n/G)}) \right) > \frac{\varepsilon}{\sqrt{2}} \right)
\rightarrow 0,
\]
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 E.1 where the convergence is uniformly in \( n \) 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 (2.9) implies (5.2) for fixed \( \alpha \) resp. \( c_{n} \). ■

Proof of Theorem 2.3 First, we obtain
\[
P \left( \max_{0 \leq k \leq n - G} \frac{1}{G} \sum_{h=0}^{\Lambda_n} \omega(h/\Lambda_n) \sum_{i=k+1}^{k+G-h} (\varepsilon_i \varepsilon_{i+h} - \gamma(h)) > \epsilon \right)
\leq \sum_{k=0}^{n-G} P \left( \sum_{h=0}^{\Lambda_n} \omega(h/\Lambda_n) \sum_{i=k+1}^{k+G-h} (\varepsilon_i \varepsilon_{i+h} - \gamma(h)) > G \epsilon \right)
\leq \frac{n}{G^2 \epsilon^2} \sum_{h=0}^{\Lambda_n} \sum_{b=0}^{G-h} \sum_{i=1}^{G-h} \sum_{j=1}^{G-b} E \left( (\varepsilon_i \varepsilon_{i+h} - \gamma(h)) (\varepsilon_j \varepsilon_{j+b} - \gamma(b)) \right)
\leq \frac{n}{G^2 \epsilon^2} \sum_{h=0}^{\Lambda_n} \sum_{b=0}^{G-h} \sum_{i=1}^{G-h} \sum_{j=1}^{G-b} |\nu(h, j - i, j - i + b)| \leq C \frac{n\Lambda_n^2}{G^2 \epsilon^2}
\]
by assumption (2.11). Consequently,
\[
\max_{G \leq k \leq n - G} \left| \frac{1}{G} \sum_{h=1}^{G-h} \omega(h/\Lambda_n) \sum_{i=k+1}^{k+G-h} (\varepsilon_i \varepsilon_{i+h} - \gamma(h)) \right| = O_P \left( \frac{n^{1/2}\Lambda_n}{G} \right).
\]

From Assumption E.2 and a Hájek-Rényi-inequality as in Lemma 3.1 for the Wiener process we get
\[
\max_{1 \leq k \leq n} \left| \sum_{i=1}^{k} \varepsilon_i \right| = O_P(\sqrt{n}).
\]
Since
\[
\max_{G \leq k \leq n - G} |\sigma_{k,n}^2 - \sigma^2| \\
\leq \max_{0 \leq k \leq n - G} \frac{1}{G} \left( \sum_{i=k+1}^{k+G} (\varepsilon_i^2 - \sigma^2) + \max_{0 \leq k \leq n - G} \frac{1}{G} \sum_{i=k+1}^{k+G} \varepsilon_i^2 \right)
\]
assertion a) follows from (5.3) (with $$\Lambda_n = 1$$) and (5.4).

Similarly,
\[
\max_{G \leq k \leq n - G} |\sigma_{k,n}^2 - \tau^2| = \max_{G \leq k \leq n - G} \left| \sigma_{k,n}^2 + 2 \sum_{h=1}^{\Lambda_n} \omega(h/\Lambda_n) \gamma_k(h) - \sigma^2 - 2 \sum_{h>0} \gamma(h) \right|
\]
\[
\leq \max_{G \leq k \leq n - G} |\sigma_{k,n}^2 - \sigma^2| \\
+ \max_{G \leq k \leq n - G} \left| \sum_{h=1}^{\Lambda_n} \omega(h/\Lambda_n) \frac{1}{G} \sum_{i=k-G+1}^{k-h} (\varepsilon_i - \tau_{k-G+1,k})(\varepsilon_{i+h} - \tau_{k-G+1,k}) - \sum_{h>0} \gamma(h) \right|
\]
\[
+ \max_{G \leq k \leq n - G} \left| \sum_{h=1}^{\Lambda_n} \omega(h/\Lambda_n) \frac{1}{G} \sum_{i=k+1}^{k+G-h} (\varepsilon_i - \tau_{k+1,k+G})(\varepsilon_{i+h} - \tau_{k+1,k+G}) - \sum_{h>0} \gamma(h) \right|
\]
\[
\leq O_P \left( \frac{\Lambda_n^{1/2}}{G} \right) + \sup_{h \in \mathbb{Z}} w(h/\Lambda_n) - 1 |\gamma(h)|,
\]
where we made repeated use of (5.4). This completes the proof of b). □

**Proof of Theorem 2.4** By assumption it holds
\[
X_i - \overline{X}_{k-G+1,k} = \varepsilon_i - \tau_{k-G+1,k} + O_P(1) \tag{5.5}
\]
where the rate is uniform in $$i, k, G$$. Hence by $$(a + b)^2 \leq 2a^2 + 2b^2$$ we get
\[
\sup_{G \leq k \leq n - G} \sigma_{k,n}^2 \leq \sup_{G \leq k \leq n - G} \frac{1}{G} \left( \sum_{i=k-G+1}^{k} (\varepsilon_i - \tau_{k-G+1,k})^2 + \sum_{i=k+1}^{k+G} (\varepsilon_i - \tau_{k+1,k+G})^2 \right) + O_P(1),
\]
implying a) by Theorem 2.3. Similarly, by (5.5) and (5.4) we get
\[
\frac{1}{G} \sum_{i=k+1}^{k+G-h} (X_i - \overline{X}_{k+1,k+G})(X_{i+h} - \overline{X}_{k+1,k+G})
\]
\[
= \frac{1}{G} \sum_{i=k+1}^{k+G-h} (\varepsilon_i - \tau_{k+1,k+G} + O_P(1)) (\varepsilon_{i+h} - \tau_{k+1,k+G} + O_P(1))
\]
\[
= \frac{1}{G} \sum_{i=k+1}^{k+G-h} (\varepsilon_i - \tau_{k+1,k+G})(\varepsilon_{i+h} - \tau_{k+1,k+G}) + O_P(1) + O_P \left( \sqrt{\frac{n}{G}} \right),
\]
where the rates are uniform in $$k, h, G$$. With Theorem 2.3 we get
\[
\max_{G \leq k \leq n - G} \frac{\sigma_{k,n}^2}{h} = O_P \left( \frac{\Lambda_n^{1/2}}{G} \right) + O_P(\Lambda_n) = O_P(\Lambda_n)
\]
concluding b). □
5 Proofs

5.2 Proofs of Section 3

Proof of Theorem 3.1 First note that (2.9) and (3.4) imply (5.2) so that by Lemma 5.1 it holds \( P(S_n) \rightarrow 1 \) with
\[
S_n = \left\{ \max_{|k-k_j| \geq \ell} \frac{|T_{k,n}(G)|}{\tau_{k,n}} < D_n(G, \alpha_n), 0 \leq |k-k_j| < (1-\varepsilon) G, j = 1, \ldots, q_n \right\}
\]
and by Theorem B.3. in [Kirch, 2006]. Analogous arguments show b) on noting that Assumption E.1 yields a) by Theorem 3.2 as well as 3.3.

Proof of Lemma 5.1 By the triangular inequality and the monotony of the \( \{c_k\} \) we get
\[
\max_{\ell \leq k \leq u} c_k \sum_{j=m+1}^{m+k} \varepsilon_j \leq c_\ell \sum_{j=m+1}^{m+\ell} \varepsilon_j + \max_{\ell \leq k \leq u} c_k \sum_{j=m+\ell+1}^{m+k} \varepsilon_j,
\]
hence
\[
P \left( \max_{\ell \leq k \leq u} \left| \sum_{j=m+1}^{m+k} \varepsilon_j \right| > \delta \right) \leq P \left( \left| \sum_{j=m+1}^{m+\ell} \varepsilon_j \right| > \frac{\delta}{2} \right) + P \left( \left| \sum_{j=m+\ell+1}^{m+k} \varepsilon_j \right| > \frac{\delta}{2} \right).
\]
An index shift in connection with the Chebyshev inequality and E.4 yields a) by Theorem B.3. in [Kirch, 2006]. Analogous arguments show b) on noting that Assumption E.3 also holds for the process in reversed time \( \{\varepsilon_{-i}\} \). ■

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 E.7 and E.8 Then it holds for any \( \beta > 0, \xi > 0 \) on \( 2G \leq k \leq n - 2G \)
\[
(a) \ P \left( \max_{k-j \leq k \leq k - \xi} \left| T_{k,n}(G; \varepsilon_1, \ldots, \varepsilon_n) - T_{k,n}(G; \varepsilon_1, \ldots, \varepsilon_n) \right| > \beta \right) = O \left( (\beta^2 G\xi)^{-\gamma/2} \right),
\]
\[
(b) \ P \left( \max_{k-j \leq k \leq k - \xi} \left| T_{k,n}(G; \varepsilon_1, \ldots, \varepsilon_n) - T_{k,n}(G; \varepsilon_1, \ldots, \varepsilon_n) \right| > \beta \right) = O \left( \beta^{-\gamma} \left( \frac{u}{G} \right)^{\gamma/2} \right),
\]
\[
(c) \ P \left( \max_{k-j \leq k \leq k - \xi} \left| T_{k,n}(G; \varepsilon_1, \ldots, \varepsilon_n) + T_{k,n}(G; \varepsilon_1, \ldots, \varepsilon_n) \right| > \beta \right) = O \left( \beta^{-\gamma} \right),
\]
where the constants only depend on \( \hat{C} \) and \( \gamma \).

Proof of Lemma 5.2 For \( G \leq k \leq G \leq k \leq G \) some straightforward calculations (confer Mulisal, 2013 (6.13)) give
\[
T_{k,n}(G; \varepsilon_1, \ldots, \varepsilon_n) - T_{k,n}(G; \varepsilon_1, \ldots, \varepsilon_n) = \frac{1}{\sqrt{2G}} \left( \sum_{i=k+G+1}^{k+G} \varepsilon_i + \sum_{i=k-G+1}^{k-G} \varepsilon_i - 2 \sum_{i=k+1}^{k} \varepsilon_i \right) \tag{5.6}
\]
Hence
\[
P\left(\max_{k_j - G \leq k_j - \xi} \left| T_{k_j,n}(G; \varepsilon_1, \ldots, \varepsilon_n) - T_{k,n}(G; \varepsilon_1, \ldots, \varepsilon_n) \right| > \beta \right) \
\leq P\left(\max_{k_j - G \leq k_j - \xi} \left| \sum_{i=k+1}^{k+G} \varepsilon_i \right| \geq \frac{\beta \sqrt{2G}}{3} \right) + P\left(\max_{k_j - G \leq k_j - \xi} \left| \sum_{i=k}^{k} \varepsilon_i \right| \geq \frac{\beta \sqrt{2G}}{3} \right)
\]
\[+ P\left(\max_{k_j - G \leq k_j - \xi} \left| \sum_{i=k+1}^{k_j} \varepsilon_i \right| \geq \frac{\beta \sqrt{2G}}{3} \right).
\]

By Lemma 3.1 and the independence of \(k_j\) and \(\varepsilon_1, \ldots, \varepsilon_n\) it follows for the first summand, where the others can be dealt with analogously,
\[
P\left(\max_{k_j - G \leq k_j - \xi} \left| \sum_{i=k+1}^{k+G} \varepsilon_i \right| \geq \frac{\beta \sqrt{2G}}{3} \right) \
\leq \tilde{C} \beta^{-2G} G^{-\gamma/2} \left(\varepsilon^{-\gamma/2} + \sum_{k = \xi+1}^{G} k^{-\gamma/2-1} \right)
\]
\[= O\left(\left(\beta^2 G \xi\right)^{-\gamma/2}\right),
\]
where \(O(1)\) only depends on \(\tilde{C}\) and \(\gamma\). The proof of b) is analogous (with \(c_k = 1\) in Lemma 3.1). Since
\[
T_{k_j,n}(G; \varepsilon_1, \ldots, \varepsilon_n) + T_{k,n}(G; \varepsilon_1, \ldots, \varepsilon_n)
= (T_{k,n}(G; \varepsilon_1, \ldots, \varepsilon_n) - T_{k_j,n}(G; \varepsilon_1, \ldots, \varepsilon_n)) + 2T_{k_j,n}(G; \varepsilon_1, \ldots, \varepsilon_n),
\]
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 E.3.

**Proof of Theorem 3.2.** We will prove
\[
P\left(\hat{k}_j > k_j + \xi_n\right) = O(1) \delta_n^{-\gamma} \left(\varepsilon^{-\gamma} + \hat{G}^{-\gamma}\right) + o(1),
\]
\[
P\left(\hat{k}_j < k_j - \xi_n\right) = O(1) \delta_n^{-\gamma} \left(\varepsilon^{-\gamma} + \hat{G}^{-\gamma}\right) + o(1),
\]
where we discuss the second assertion in detail, the first one follows analogously (where an analogous version of Lemma 5.3 is needed). Consider the set
\[
M_n = \{q_n = q_n, \max_{1 \leq j \leq q_n} |\hat{k}_j - k_j| < G \min_{G \leq k \leq n - G} \hat{\tau}_{k,n} > 0 \}
\]
\[\cap \left\{ \min_{j=1, \ldots, q_n} |d_j| \geq \delta_n, \min_{j=1, \ldots, q_n} |k_{j+1} - k_j| > 2\hat{G} \right\}, \quad (5.7)
\]
which does not depend on \(j\). Define \(\tilde{w}_j := \min(w_j, k_j + \hat{G} - 1)\) and \(\tilde{v}_j := \max(v_j, k_j - G + 1)\), then it holds on \(M_n\) for all \(j = 1, \ldots, q_n\)
\[
\hat{k}_j = \arg\max_{\tilde{v}_j \leq k \leq \tilde{w}_j} \frac{|T_{k,n}(G)|}{\hat{\tau}_{k,n}}.
\]
Next, note that
\[
\arg\max_{\tilde{v}_j \leq k \leq \tilde{w}_j} \frac{|T_{k,n}(G)|}{\hat{\tau}_{k,n}} = \arg\max_{\tilde{v}_j \leq k \leq \tilde{w}_j} \frac{V_{k,n}^{(j)}(G)}{\hat{\tau}_{k,n}}, \quad V_{k,n}^{(j)}(G) = (T_{k,n}(G))^2 - (T_{k,n}(G))^2.
\]
Since on $M_n$ we get $D_{k,n} \leq P_{k,n} \leq 1$, we have

$$V_{k,n}^{(j)}(G) = \arg\max_{\hat{v}_j \leq k \leq k_j - \xi_n} \frac{V_{k,n}^{(j)}(G)}{\tau_{k,n}}$$

Hence,

$$P \left( \arg\max_{\hat{v}_j \leq k \leq k_j - \xi_n} \frac{V_{k,n}^{(j)}(G)}{\tau_{k,n}} < k_j - \xi_n \right)$$

$$= P \left( \max_{\hat{v}_j \leq k \leq k_j - \xi_n} \frac{V_{k,n}^{(j)}(G)}{\tau_{k,n}} \geq \arg\max_{k_j - \xi_n \leq k \leq \hat{v}_j} \frac{V_{k,n}^{(j)}(G)}{\tau_{k,n}} \right)$$

$$\leq P \left( \max_{\hat{v}_j \leq k \leq k_j - \xi_n} \frac{V_{k,n}^{(j)}(G)}{\tau_{k,n}} \geq 0 \right)$$

$$\leq P \left( \max_{\hat{v}_j \leq k \leq k_j - \xi_n} V_{k,n}^{(j)}(G) \geq 0 \right) + P \left( \min_{\hat{v}_j \leq k \leq \hat{v}_n - \xi_n} \tau_{k,n} \leq 0 \right).$$

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

$$V_{k,n}^{(j)}(G) = (T_{k,n}(G))^2 - (T_{k_j,n}(G))^2 = (T_{k,n}(G) - T_{k_j,n}(G))(T_{k,n}(G) + T_{k_j,n}(G))$$

$$= - \left( (T_{k_j,n}(G;\hat{v}_1,\ldots,\hat{v}_n) - T_{k,n}(G;\hat{v}_1,\ldots,\hat{v}_n)) + (2G)^{-1/2}(k_j - k)d_j \right)$$

$$\cdot \left( (T_{k,n}(G;\hat{v}_1,\ldots,\hat{v}_n) + T_{k,n}(G;\hat{v}_1,\ldots,\hat{v}_n)) + (2G)^{-1/2}(k + 2G - k_j)d_j \right)$$

$$=: -(A_1(k,n) + D_1(k,n))(A_2(k,n) + D_2(k,n)).$$

(5.8)

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

$$P \left( \max_{k_j - G \leq k \leq k_j - \xi_n} V_{k,n}^{(j)}(G) \geq 0, M_n \right)$$

$$\geq P \left( \max_{k_j - G \leq k \leq k_j - \xi_n} \frac{A_1(k,n)}{D_1(k,n)} \frac{A_2(k,n)}{D_2(k,n)} + \frac{A_1(k,n)}{D_1(k,n)} \frac{A_2(k,n)}{D_2(k,n)} \geq 0, M_n \right)$$

$$\leq P \left( \max_{k_j - G \leq k \leq k_j - \xi_n} \left| \frac{A_1(k,n)}{D_1(k,n)} \frac{A_2(k,n)}{D_2(k,n)} \right| \geq 1, M_n \right)$$

$$\leq P \left( \max_{k_j - G \leq k \leq k_j - \xi_n} \left| \frac{A_1(k,n)}{k_j - k} \right| \geq \frac{\delta_n}{\sqrt{2G}} \right) + P \left( \max_{k_j - G \leq k \leq k_j - \xi_n} \left| \frac{A_2(k,n)}{k_j - k} \right| \geq \frac{\delta_n}{\sqrt{6G}} \right)$$

$$\leq O(1) \delta_n^{-\gamma} \left( \xi_n^{-\hat{z}} + G^{-\hat{z}} \right),$$

where the last line follows from Lemma 5.2 and $O(1)$ does not depend on $j$. This concludes the proof of a). Similarly, on the set $\tilde{M}_n = M_n \cap \{ g_n \leq \gamma_n \}$

(5.9)

we get

$$P \left( \max_{j=1,\ldots,\min(\hat{q}_n,\gamma_n)} |k_j - k_j| > \xi_n, \tilde{M}_n \right) \leq \sum_{j=1}^{\gamma_n} P \left( |\hat{k}_j - k_j| > \xi_n, \tilde{M}_n \right)$$

$$= O(1) \gamma_n \delta_n^{-\gamma} \left( \xi_n^{-\hat{z}} + G^{-\hat{z}} \right),$$

22
where the last line follows from the proof of a). Since \( P(\tilde{M}_n) \to 1 \), the proof of b) is complete. ■

**Proof of Theorem 3.3** The decomposition \((5.8)\) yields for \( k < k_j \)

\[
V_{k,n}^{(2)}(G) = -D_1(k,n)D_2(k,n) - A_1(k,n)D_2(k,n) - A_2(k,n)D_1(k,n) - A_1(k,n)A_2(k,n).
\]

By \( E.1 \) 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

\[
\max_{1 \leq k_j - k \leq c \sqrt{d_j^2}} |A_1(k,n)| = o_P(1), \quad \max_{1 \leq k_j - k \leq c \sqrt{d_j^2}} |A_2(k,n)| = o_P(1).
\]

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

\[
V_{k,n}^{(2)}(G) = - \frac{1}{2G} |k_j - k| (2G - |k_j - k|) d_j^2
\]

\[
- (T_{k,n}(G; \varepsilon_1, \ldots, \varepsilon_n) - T_{k,n}(G; \varepsilon_1, \ldots, \varepsilon_n))(2G - |k_j - k|) \frac{1}{\sqrt{2G}} d_j + o_P(1)
\]

\[
= - |k_j - k| d_j^2 - d_j \left( \sum_{i=k+G+1}^{k+G} \varepsilon_i + \sum_{i=k-G+1}^{-G+1} \varepsilon_i - 2 \sum_{i=k+1}^{k} \varepsilon_i \right) + o_P(1).
\]

By stationarity and \( E.1 \) we get

\[
\left\{ d_j \left( \sum_{i=k+G+1}^{k+G} \varepsilon_i + \sum_{i=k-G+1}^{-G+1} \varepsilon_i - 2 \sum_{i=k+1}^{k} \varepsilon_i \right) : k = k_j - 1, \ldots, k_j - c \sigma^2 d_j^{-2} \right\} 
\]

\[
\overset{P}{=} \left\{ U_n(l) = d_j \left( \sum_{i=-l+G+1}^{G} \varepsilon_i + \sum_{i=-l-G+1}^{-G+1} \varepsilon_i - 2 \sum_{i=-l+1}^{0} \varepsilon_i \right) : 1 \leq l \leq c \sigma^2 d_j^{-2} \right\}.
\]

Note that by the assumption on \( d_j \) and \( E.1 \) the three summands are asymptotically independent. Hence by \( d_j \overset{P}{\to} 0 \) and \( E.1 \) the functional central limit theorem implies

\[
\left\{ \frac{U_n(s\sigma^2 d_j^{-2})}{\tau^2} : 0 \leq s \leq c \right\} \overset{D[0,1]}{\to} \{ \sqrt{6} W(s), 0 \leq s \leq c \},
\]

because \( \{ W_1(s) + W_2(s) - 2W_3(s) \} \overset{D}{=} \{ -\sqrt{6} W(s) \} \) 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 \( k_j \geq k \), which implies by \( P(M_n) \to 1 \) for \( -c \leq x \leq c \)

\[
P \left( -c \leq d_j \frac{k_j 1_{(j \leq \hat{o}_n)} - k_j}{\tau^2} \leq x \right)
\]

\[
\to P \left( \max_{-c \leq s \leq x} \left( -|s| - \sqrt{6} W(s) \right) \geq \max_{x \leq s \leq c} \left( -|s| - \sqrt{6} W(s) \right) \right)
\]

\[
= P \left( -c \leq \arg \max_{-c \leq s \leq c} \left( W(s) - |s|/\sqrt{6} \right) \leq x \right).
\]

By Theorem 3.2 and Remark 3.2 a) we get

\[
P \left( d_j^2 \frac{k_j - k_j}{\tau^2} \leq c \right) \leq \left( c^{-\gamma/2} + G^{-\gamma/2} \right) O(1) + o(1)
\]
uniformly in $n$, which becomes arbitrarily small for $c$ large enough. Hence, letting $c \to \infty$ gives assertion a). Assertion b) follows because on

$$\{q_n = q\} \cap \bigcap_{j=1}^{q_n} \left\{\sum_{j=1}^{q_n} \frac{|k_j - k_j^*|}{\tau^2} \leq c\right\}$$

the change point estimators $\hat{k}_j$, $j = 1, \ldots, n$, are asymptotically independent (conditionally on $k_j, d_j$, $j = 1, \ldots, q_n$) by Assumption E.4. Because the above convergence also holds conditionally, this implies b) by using the dominated convergence theorem in the very last step.

**References**


References


References


