Annals of the Institute of Statistical Mathematics (2024) 76:393–421
https://doi.org/10.1007/s10463-023-00892-4

Data segmentation for time series based on a general moving sum approach

Claudia Kirch[1,2] · **Kerstin Reckruehm**[1]

Received: 19 July 2022 / Revised: 22 September 2023 / Accepted: 27 September 2023 /
Published online: 14 March 2024
© The Institute of Statistical Mathematics, Tokyo 2024

Abstract
We consider the multiple change point problem in a general framework based on estimating equations. This extends classical sample mean-based methodology to include robust methods but also different types of changes such as changes in linear regression or changes in count data including Poisson autoregressive time series. In this framework, we derive a general theory proving consistency for the number of change points and rates of convergence for the estimators of the locations of the change points. More precisely, two different types of MOSUM (moving sum) statistics are considered: A MOSUM-Wald statistic based on differences of local estimators and a MOSUM-score statistic based on a global inspection parameter. The latter is usually computationally less involved in particular in nonlinear problems where no closed form of the estimator is known such that numerical methods are required. Finally, we evaluate the methodology by some simulations as well as using geophysical well-log data.

Keywords Multiple change points · M-estimators · Estimating function · Robust statistics · Poisson autoregression · Linear regression · Median

1 Introduction

Data segmentation or multiple change point estimation is a current topic in statistics and machine learning and comprises problems in a wide range of fields like finance, quality control, medicine or climate. For example, Braun et al. (2000) apply a technique of multiple change point detection on DNA sequences. Aggarwal et al. (1999) focus on finding structural breaks in the volatility of stock market returns,

✉ Claudia Kirch
claudia.kirch@ovgu.de

[1] Department of Mathematics, Institute of Mathematical Stochastics, Otto-von-Guericke University, Universitätsplatz 2, 39106 Magdeburg, Germany

[2] Center for Behavioral Brain Science (CBBS), Magdeburg, Germany

while Killick et al. (2010) and Killick et al. (2012) give an interesting application to oceanography by detecting changes in the variance of time series for wave heights.

Change point analysis has a long tradition in statistics dating back as long as Page (1954). From the beginning, research on change points has been focused on two different directions: Sequential or online change point analysis on the one hand (see, e.g., the recent books by Tartakovsky et al. (2014) and Tartakovsky (2019)) and a-posteriori or offline change point analysis (see, e.g., the books by Csörgö and Horváth (1997) or Chen and Gupta (2012)) on the other hand, where this work belongs to the second type. The early a-posteriori literature focused on at-most-one-change testing and the corresponding estimator for a single change point, which is still an area of interest for changes in more complex data structures, see, e.g., the survey articles by Aue and Horváth (2013); Horváth and Rice (2014); Cho and Kirch (2021). At the same time, the theoretic investigation of the multiple change situation with all its additional challenges has become increasingly popular, where most papers deal with the detection of multiple changes in the mean; see for example the recent survey articles by Fearnhead and Rigaill (2020), Cho and Kirch (2021) or Truong et al. (2020). There are many different approaches to the problem ranging from Bayesian methods (e.g., Fearnhead (2006) or Mohammad-Djafari and Féron (2006)), global optimisation methods based on suitable information criteria (for a detailed review see Cho and Kirch (2021), Section 3.1) to methods based on local testing procedures (for a detailed review see Cho and Kirch (2021), Section 3.2). The latter includes several variations of binary segmentation algorithms (e.g., Vostrikova (1981) or Fryzlewicz (2014) for the usual mean changes and Korkas and Fryzlewicz (2017) for changes in the second-order structure of time series) but also moving sum (MOSUM) procedures. In this paper, we adopt the latter approach.

MOSUM statistics for multiple changes in the mean were first mathematically analyzed in detail by Eichinger and Kirch (2018). Corresponding tests had previously been considered by Bauer and Hackl (1980), Hušková (1990), Chu et al. (1995) and Hušková and Slabỳ (2001). Cho and Kirch (2022) refine the procedure by an additional post-processing step based on an information criterion. An efficient implementation of the procedure in an R-package is detailed in Meier et al. (2021). Yau and Zhao (2016) use moving sum statistics in combinatios with an information criterion for estimating the change points in a linear autoregressive setting. These methods are based on L_2-methodology for changes in the mean but do neither allow for robust statistical methodology nor are they suitable for more complex changes such as changes in (non-)linear (auto-)regression or changes in count time series. Extensions to continuous-time processes in particular renewal processes are proposed by Messer et al. (2014) as well as Kirch and Klein (2023).

In statistics, a successful general theoretic framework that can be used for a variety of problems is based on so-called estimating equations and also related to M-estimation. In change point analysis, for the at-most-one-change test and corresponding change point estimator, this was first considered in the context of robust methodology e.g., Huskova (2013) but later seen to include many different types of changes in time series like nonlinear (auto-)regressive time series based on neural

network approximations but also changes in different count time series, e.g., Kirch and Kamgaing (2016). Later this framework was also considered in the context of sequential change point analysis by Kirch and Kamgaing (2015); Kirch and Weber (2018) but so far it has not been used for multiple change detection.

Such an analysis in combination with single-bandwidth MOSUM procedures as in Eichinger and Kirch (2018) is the aim of the present paper thus providing the theoretical background including statistical guarantees for a variety of new data segmentation procedures for various important examples including more robust methods for mean changes and methods for various time series models. This theoretic framework is different from the one obtained in Kirch and Klein (2023) and with the exception of the multiple mean change scenario does not include common examples. In particular, in Eichinger and Kirch (2018) as well as Kirch and Klein (2023) methodology is directly based on increments of an observed process (in the mean change this corresponds to the partial sum process). In the context of general estimating functions, the corresponding moving sums of the estimating functions do not only depend on the observed data but also the unknown parameter that can change with time. Also, unlike in the above two examples, there is more than one possible extension of the MOSUM methodology related to Wald- and score-type procedures. In this paper, we introduce and analyze both versions from a theoretical perspective where the same (optimal) consistency results for the change point estimators are obtained as in the original mean change situation. It turns out that both methods have different strengths and weaknesses which will play an important role if used as candidate generating methods in combination with post-processing methods such as discussed for L_2 procedures for mean changes by Cho and Kirch (2022), see Sect. 3.3. At the same, the theoretical findings of this paper provide the basis for corresponding results of the post-processing methodology as the post-processing essentially inherits the favorable properties of the candidate generating mechanism in the mean change situation, see Cho and Kirch (2022).

The paper is organized as follows: In Sect. 2 we explain in detail the two data segmentation algorithms that we propose. The discussion in this paper occurs in a general framework based on estimating functions where examples are given in Subsection 2.3. In Sect. 3, we prove consistency of the data segmentation algorithms including localization rates for one of the procedures. Due to the general framework the consistency results are obtained based on some high-level assumptions. In Sect. 4, we prove these high-level assumptions for sufficiently smooth estimating functions under standard moment conditions. In Sect. 5, the empirical performance is investigated by means of a simulation study and using a geophysical well-data set before we give some conclusions in Sect. 6. The proofs and some additional simulations can be found in the supplementary material.

2 Data segmentation based on moving sum statistics

2.1 MOSUM-Wald and MOSUM-score statistics

We consider the following general setting

$$X_t = \sum_{j=1}^{q+1} X_t^{(j)} 1_{\{k_{j-1,n} < t \leq k_{j,n}\}}, \quad k_{0,n} = 0,\ k_{q+1,n} = n. \tag{1}$$

The sequences $\{X_t^{(j)} : t \geq 1\}, j = 1, \ldots, q+1$, are assumed to be stationary, where consecutive sequences are distributionally different. Then, q denotes the number of structural breaks and $k_{1,n}, \ldots, k_{q,n}$ denote the change points. For simplicity we assume that q is fixed but all arguments are valid for a sequence q_n as long as q_n is bounded. Under some additional assumptions, it is also possible to relax the boundedness assumption on the number of change points (see e.g., Kirch and Klein (2023) or Section E.3 in Cho and Kirch (2022)).

While the construction of the below statistics is based on a given parametric model, we do not require the observed time series to follow that model. Instead, we explicitly allow for model misspecification in the theoretic analysis showing that under misspecification changes are detected if consecutive time series have different best approximating model parameters (in the sense of Assumption 2).

The test statistic is related to estimators of the model parameters based on estimating functions $\boldsymbol{H}$ also known as M-estimators. These are obtained as the solution $\widehat{\boldsymbol{\theta}}_{a,b}$ of the estimating equation system $\sum_{i=a}^{b} \boldsymbol{H}(\mathbb{X}_i, \boldsymbol{\theta}) \stackrel{!}{=} \mathbf{0}$ for a suitable choice of $\mathbb{X}_i$, which can be equal to the observations, a tuple of response and (stationary) explanatory variables or (for autoregressive models) can include lagged observations; see Sect. 2.3 for some examples. The estimating function $\boldsymbol{H}$ is vector-valued for the estimation of multidimensional parameter vectors.

As we adopt a general framework in this paper and explicitly allow for misspecification, we need to make some high-level assumptions on the underlying time series. We show their validity under some moment conditions for sufficiently smooth estimating functions in Sect. 4.

Assumption 1

(a) For each segment j, let $\{\mathbb{X}_t^{(j)} : t \geq 1\}$ be stationary.

(b) For a given $\boldsymbol{\theta}$ specified below let $\boldsymbol{S}_j(k, \boldsymbol{\theta}) = \sum_{i=1}^{k} \boldsymbol{H}(\mathbb{X}_i^{(j)}, \boldsymbol{\theta})$ fulfill a strong invariance principle for all $j = 1, \ldots, q+1$, i.e., possibly after changing the probability space there exists some $\nu > 0$, a p-dimensional standard Wiener process $\{\boldsymbol{W}(k) : k \geq 0\}$ and a symmetric positive definite long-run covariance matrix $\boldsymbol{\Sigma}_{(j)}(\boldsymbol{\theta})$ such that as $k \to \infty$

$$\left\| \boldsymbol{S}_j(k, \boldsymbol{\theta}) - E(\boldsymbol{S}_j(k, \boldsymbol{\theta})) - \boldsymbol{\Sigma}_{(j)}(\boldsymbol{\theta})^{1/2} \boldsymbol{W}(k) \right\| = O(k^{1/(2+\nu)}) \ \ a.s.$$

In the literature, invariance principles as in (b) have been derived for many time series – compare also Theorem 5 below, where typically the parameter ν depends on the number of existing moments.

The following assumption is a Bahadur representation for M-estimators as for example derived by He and Shao (1996) but uniformly in k. Effectively, this is a linearisation of the estimator which is typically used to derive asymptotic normality. In our case, the representation will also be used to make a connection between the MOSUM-Wald and the MOSUM-score statistics in the proof.

Assumption 2 There exists a regular matrix $\boldsymbol{V}_{(j)}$ for any $j = 1, \ldots, q+1$ such that

$$\max_{k_{j-1,n}<k\le k_{j,n}-G} \left\| \sqrt{\frac{G}{2}} \boldsymbol{V}_{(j)} \left(\boldsymbol{\theta}_j - \widehat{\boldsymbol{\theta}}_{k+1,k+G} \right) - \frac{1}{\sqrt{2G}} \sum_{i=k+1}^{k+G} \boldsymbol{H}(\mathbb{X}_i^{(j)}, \boldsymbol{\theta}_j) \right\| = o_P\left((\log(n/G))^{-1/2} \right),$$

where $\boldsymbol{\theta}_j$ fulfills $E\left(\boldsymbol{H}(\mathbb{X}_i^{(j)}, \boldsymbol{\theta}_j) \right) = \mathbf{0}$ (identifiably unique).

If the parametric assumption underlying the M-estimator is correct, then $\boldsymbol{\theta}_j$ is the true underlying parameter for the jth stationary sequence. Otherwise, this is the best approximating parameter in the sense induced by the estimating function. Furthermore, in case of differentiable estimating functions, it usually holds $\boldsymbol{V}_{(j)} = E\left(\nabla \boldsymbol{H}(\mathbb{X}_1^{(j)}, \boldsymbol{\theta}_j)^T \right)$; see Regularity Condition 2 (b).

Moreover, we define

$$\begin{aligned} &\boldsymbol{V}_k = \boldsymbol{V}_{(j)},\ \boldsymbol{\Sigma}_k(\boldsymbol{\theta}) = \boldsymbol{\Sigma}_{(j)}(\boldsymbol{\theta}),\ \boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_{(j)}(\boldsymbol{\theta}_j) \quad \text{for } k_{j-1,n} < k \le k_{j,n}, \\ &\boldsymbol{\Gamma}_k = \boldsymbol{V}_k^{-1} \boldsymbol{\Sigma}_k \left(\boldsymbol{V}_k^{-1} \right)^T, \qquad \boldsymbol{\Gamma}_{(j)} = \boldsymbol{\Gamma}_{k_{j,n}}, \end{aligned} \tag{2}$$

where $\boldsymbol{\Gamma}_k$ is the asymptotic (long-run) covariance matrix for the estimators from the jth stationary segment under the above assumptions.

A first intuitive approach to the multiple change problem is based on the following **MOSUM-Wald statistic** that uses weighted differences of moving estimators $\widehat{\boldsymbol{\theta}}_{k+1,k+G}$ and $\widehat{\boldsymbol{\theta}}_{k-G+1,k}$ of the unknown quantity θ based on the stretch of data $X_{k+1}, \ldots, X_{k+G}$, respectively, $X_{k-G+1}, \ldots, X_k$. Similarly to the Mahalanobis distance, the weighting is done with the asymptotic (long-run) covariance matrix, so that we obtain for $k = G, \ldots, n-G$

$$T_{k,n}^{(1)}(G) = T_{k,n}^{(1)}(G; \boldsymbol{\Gamma}_k) = \frac{\sqrt{G}}{\sqrt{2}} \left\| \boldsymbol{\Gamma}_k^{-1/2} \left(\widehat{\boldsymbol{\theta}}_{k+1,k+G} - \widehat{\boldsymbol{\theta}}_{k-G+1,k} \right) \right\|, \tag{3}$$

where the bandwidth $G = G_n$ is a tuning parameter that determines the length of the moving window, and $\|\cdot\|$ denotes the Euclidian norm.

In this paper, we consider bandwidths G that are of smaller order than the sample length n. The procedure detects changes that are further apart than $2G$ such

that sublinear changes, whose distance diverges strictly slower than the sample size, can be detected.

Assumption 3

(a) For $\nu > 0$ (specified in Assumption 1 (b)) let

$$\frac{n}{G} \to \infty \text{ and } \frac{n^{\frac{2}{2+\nu}} \log(n)}{G} \to 0 \text{ for } n \to \infty.$$

(b) The minimal distance between two neighboring change points is asymptotically larger than $2G$ in the sense of

$$\liminf_{n\to\infty} \min_{j=1,\dots,q+1} (k_{j,n} - k_{j-1,n})/G > 2.$$

Close to a change point the above Wald statistic (3) can be expected to be large such that it can be used to find changes in the parameter vector $\boldsymbol{\theta}$. However, using the Wald statistic has one major drawback: Calculating two estimates for each time point, i.e., $2(n-2G)$ in total, can be computationally challenging in situations where numerical methods need to be applied as, e.g., in nonlinear models such as Poisson autoregressive models. Furthermore, in situations with many (almost) zeroes (corresponding to many local optima) such as in a neural network autoregressive situation as in Kirch and Kamgaing (2012) estimators can be far apart even though the corresponding stochastic processes are almost identical. While this is explicitly excluded for our theoretic results by assuming identifiability, this can easily lead to problems in applications.

In order to avoid these problems and to reduce the computational complexity and corresponding numerical challenges, we consider **MOSUM-score statistics** where local parameter estimators are replaced by a global inspection parameter $\widetilde{\boldsymbol{\theta}}$, which can be fixed or an estimator based on the same data. The MOSUM-score statistic is based on the following differences between moving sums of the estimating function at the inspection parameter for $k = G, \dots, n-G$

$$\boldsymbol{M}_{\widetilde{\theta}}(k) = \sum_{i=k+1}^{k+G} \boldsymbol{H}(\mathbb{X}_i, \widetilde{\boldsymbol{\theta}}) - \sum_{i=k-G+1}^{k} \boldsymbol{H}(\mathbb{X}_i, \widetilde{\boldsymbol{\theta}}).$$

Effectively, this converts a general multiple parameter change problem to a multiple mean change problem of the transformed sequence $\{\boldsymbol{H}(\mathbb{X}_t, \widetilde{\boldsymbol{\theta}})\}_{t\geq 1}$. Consequently, a multivariate version of the (univariate) mean-MOSUM statistic investigated by Eichinger and Kirch (2018) can be applied, leading to the following MOSUM-score statistic:

$$T_{k,n}^{(2)}(G) = T_{k,n}^{(2)}(G, \widetilde{\boldsymbol{\theta}}) = T_{k,n}^{(2)}(G, \widetilde{\boldsymbol{\theta}}; \boldsymbol{\Sigma}_k(\widetilde{\boldsymbol{\theta}})) = \frac{1}{\sqrt{2G}} \left\| \boldsymbol{\Sigma}_k(\widetilde{\boldsymbol{\theta}})^{-1/2} \boldsymbol{M}_{\widetilde{\theta}}(k) \right\|. \tag{4}$$

If we use data-dependent inspection parameters $\widetilde{\boldsymbol{\theta}}_n = \widetilde{\boldsymbol{\theta}}_n(\mathbb{X}_1, \ldots, \mathbb{X}_n)$, then we need the following additional assumptions for the MOSUM-score procedure:

Assumption 4 There exists $\widetilde{\boldsymbol{\theta}}$ such that for any $j = 1, \ldots, q$

$$(i) \quad \max_{k_{j-1,n}+G \le k \le k_{j,n}-G} \left\| \boldsymbol{M}_{\widetilde{\boldsymbol{\theta}}_n}(k) - \boldsymbol{M}_{\widetilde{\boldsymbol{\theta}}}(k) \right\| = o_P\left(\sqrt{\frac{G}{\log(n/G)}} \right),$$

$$(ii) \quad \max_{|k-k_{j,n}|<G} \left\| \boldsymbol{M}_{\widetilde{\boldsymbol{\theta}}_n}(k) - \boldsymbol{M}_{\widetilde{\boldsymbol{\theta}}}(k) \right\| = o_P\left(\sqrt{G \log(n/G)} \right).$$

This assumption holds under smoothness and mixing assumptions for any $\sqrt{n}$ -consistent estimator $\widetilde{\boldsymbol{\theta}}_n$ (see Theorem 8 below). In particular, $\widetilde{\boldsymbol{\theta}}_n$ obtained from the corresponding estimating equations based on the stretch of data $X_a, \ldots, X_b$ are $\sqrt{n}$-consistent for $\widetilde{\boldsymbol{\theta}}$, the best approximating parameter in the sense of Theorem 7. While for the latter approach, some theoretical guarantees toward detectability are given in Sect. 3.2.1, from a computational perspective it might be more efficient to use different estimators (for an empirical example, see Sect. 5.1).

The latter MOSUM-score approach avoids the numerical drawbacks of the MOSUM-Wald statistic but can only detect changes for which the given inspection parameter $\widetilde{\boldsymbol{\theta}}$ results in a change in the expectation of the transformed series. This problem will be discussed in detail in Sect. 3.2.1.

In the mean change model based on moving sample means as considered by Eichinger and Kirch (2018) (see also Example 2.3.1 below), the above MOSUM-Wald and MOSUM-score statistics coincide for any inspection parameter $\widetilde{\boldsymbol{\theta}}$.

2.2 Segmentation algorithm

MOSUM statistics as introduced in the previous section have peaks close to the true change points making them particularly suitable for data segmentation. More precisely, as demonstrated in Fig. 1, the MOSUM statistic is a noisy version of the MOSUM signal, which is a piecewise linear function that is equal to zero away from the change points and has a single peak at each change point as long as Assumption 3 (b) is met. The height of the peaks depends on a combination of the magnitude of the change point and the bandwidth.

Clearly, in order to use this noisy signal for segmentation purposes, a threshold is needed to distinguish between significant local maxima that are due to a close-by change point and local maxima obtained simply from random fluctuations around zero if no change point is close by. We obtain such thresholds by globally controlling the random fluctuations of the MOSUM statistic asymptotically if no change point is present in the time series. As a consequence, the proposed procedure controls the (asymptotic) family-wise error rate at level α_n (see below) for the detected change points. In Sect. 2.4 below, we detail how to choose such a threshold $D_n(\alpha_n, G)$ based on asymptotic α_n-quantiles (of the no-change situation). These quantiles can also be considered as critical values for a corresponding uniform (across time) test procedure for the null hypothesis of no change versus the

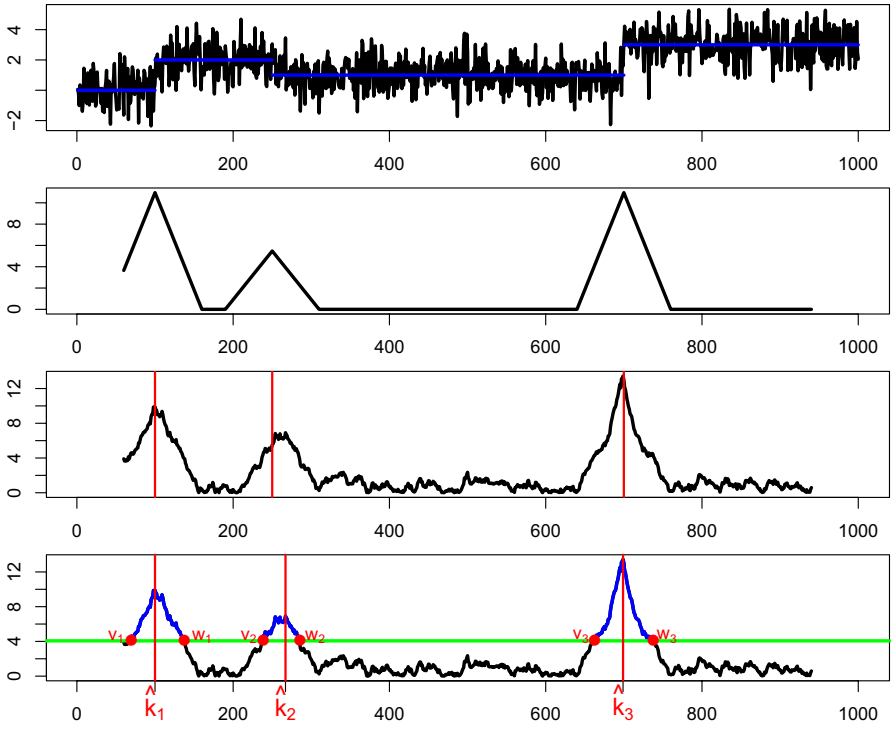


Fig. 1 Top panel: Time series with changes in the mean at time points 100, 250 and 700. Second panel: Signal of the classical MOSUM statistic for mean changes as investigated by Eichinger and Kirch (2018) with bandwidth $G = 60$. Third panel: Actual MOSUM statistic (i.e., signal plus remaining noise) for the above data set (the vertical lines indicate the true change points). Fourth panel: Outcome of the segmentation procedure as described in Sect. 2.2 in this example

alternative of one or more changes. An asymptotic threshold is reasonable given the generality of our procedure, the (nonparametric) error structure and as it also allows for model misspecification.

Based on a MOSUM-Wald or MOSUM-score statistic $T_{k,n}(G)$ with corresponding threshold $D_n(\alpha_n, G)$ we propose the following **MOSUM segmentation procedure** to determine estimators for the number and the locations of the change points. Below, we suppress the possible dependence on other tuning parameters such as the inspection parameter $\widetilde{\theta}$ for the MOSUM-score statistic:

Consider all pairs of time points $(v_{j,n}, w_{j,n})$, $v_{1,n} < w_{1,n} < v_{2,n} < w_{2,n} < \ldots$, *with*

$$\begin{aligned} &T_{k,n}(G) \geq D_n(\alpha_n, G) \quad \text{for} \quad v_{j,n} \leq k \leq w_{j,n}, \\ &T_{k,n}(G) < D_n(\alpha_n, G) \quad \text{for} \quad k = v_{j,n} - 1, w_{j,n} + 1, \\ &w_{j,n} - v_{j,n} \geq \varepsilon G \quad \text{with} \quad 0 < \varepsilon < 1/2 \quad \text{arbitrary but fixed.} \end{aligned} \tag{5}$$

We take the number of these pairs as an estimator for the number of changes:

$$\widehat{q}_n = \text{ number of pairs } (v_{j,n}, w_{j,n}).$$

Furthermore, we determine the local maxima between $v_{j,n}$ and $w_{j,n}$, $j = 1, \dots, \widehat{q}_n$, and use them as estimators for the locations of the change points:

$$\widehat{k}_{j,n} = \underset{v_{j,n} \le k \le w_{j,n}}{\arg\max}\, T_{k,n}(G).$$

The pairs $(v_{j,n}, w_{j,n})$, $j = 1, \dots, \widehat{q}_n$, give start and end points of intervals on which the statistic exceeds the threshold. For this reason, we call $[v_{j,n}, w_{j,n}]$, $j = 1, \dots, \widehat{q}_n$, intervals of exceedings or exceeding intervals.

Condition (5) avoids false positives obtained due to random fluctuation when the linear signal crosses the threshold. While the linear signal is strictly monotone when crossing the threshold, the noisy version may not be – possibly jumping beneath the threshold and then above again simply by chance.

By the below theory (see Sect. 3), we can be relatively certain (with asymptotic probability one), that the exceeding intervals contain a change point. Typically, $T_{k,n}(G)$ depends on an estimator for the long-run covariance matrix $\mathbf{\Gamma}_k$ resp. $\mathbf{\Sigma}_k$, where in many situations a local estimator (depending on the time point k) is required because the long-run covariance matrix depends on the segment. However, once the exceeding intervals have been determined, different and possibly better estimators $\widehat{\mathbf{\Psi}}_{j,n}$, $j = 1, \dots, \widehat{q}_n$, can be used to obtain the final change point estimator

$$\widehat{k}_{j,n}(\widehat{\mathbf{\Psi}}_{j,n}) = \underset{v_{j,n} \le k \le w_{j,n}}{\arg\max}\, T_{k,n}(G; \widehat{\mathbf{\Psi}}_{j,n}), \tag{6}$$

where $\widehat{\mathbf{\Psi}}_{j,n}$ is a sequence of symmetric positive definite matrices fulfilling Assumptions 9 below. For a real-valued estimating function $\mathbf{H}$, this is always equivalent to using the unscaled maximizer, i.e., $\widehat{\mathbf{\Psi}}_{j,n} = 1$. For a vector-valued estimating function it is also possible (e.g., for computational reasons) to use the identity matrix $\widehat{\mathbf{\Psi}}_{j,n} = \mathrm{Id}$ without jeopardizing the asymptotic localization rates. Indeed, the theory does not even require the $\widehat{\mathbf{\Psi}}_{j,n}$ to be consistent estimators for the long-run covariance matrix. While we use the initial estimator in the simulation study, we obtain the localization rates for the latter. A corresponding analysis for covariance estimators depending on the time point k is not possible in our general framework, but can be done for specific estimators (see, e.g., Remark 3.2 and Corollary 3.1 in Eichinger and Kirch (2018) for an example).

2.3 Examples

The framework that is used in this work covers many different situations (see, e.g., Kirch and Kamgaing (2015) and Kirch and Weber (2018) for additional examples). Here, we concentrate on the following three examples: Uni- and multivariate changes in expectation where the sample mean but also more robust estimators can be used, linear regression and Poisson autoregression.

2.3.1 Mean change model

The univariate mean change model is given by

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

where $\{\varepsilon_i : i \geq 1\}$ is a centered stationary error sequence with autocovariance function $\gamma(\cdot)$ and long-run variance $0 < \tau^2 = \sum_{h \in \mathbb{Z}} \gamma(h) < \infty$. Thus, the expected value of each stretch is given by $\mu_j \neq \mu_{j+1}$, $j = 1, \ldots, q$. The MOSUM statistic discussed in Eichinger and Kirch (2018) is based on the sample mean $\bar{X}_{1,n} = \frac{1}{n} \sum_{i=1}^{n} X_i$ which is the solution of the estimating equation $\sum_{i=1}^{n} (X_i - \mu) \stackrel{!}{=} 0$, hence $\mathbb{X}_i = X_i$.

An alternative M-estimator for the expectation for symmetric error distributions is based on the estimating function $H(X_i, \mu) = \frac{2}{\pi} \arctan(\mu - X_i)$. This estimating function leads to more robust estimators and can be seen as a smooth approximation of the sign-function that leads to the median. Henceforth, we will therefore call the corresponding estimator *median-like estimator*.

A multivariate version $\boldsymbol{X}_i = (X_{1,i}, \ldots, X_{d,i})^T$ is obtained by replacing ε_j as well as μ_j with multivariate quantities. The multivariate estimating function is given by $\boldsymbol{H}(\boldsymbol{X}_i, \boldsymbol{\mu}) = (H(X_{1,i}, \mu_1), \ldots, H(X_{d,i}, \mu_d))^T$ for $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^T$, i.e., $\mathbb{X}_i = \boldsymbol{X}_i$. Effectively, the procedure uses the corresponding univariate estimators in each component. Consequently, an estimator for the long-run covariance matrix is now required which poses a major difficulty in practice.

2.3.2 Linear regression model with structural breaks

The linear regression model with changes is given by

$$X_i = \sum_{j=1}^{q+1} \boldsymbol{Z}_i^T \boldsymbol{\beta}_j \, 1_{\{k_{j-1,n} < i \leq k_{j,n}\}} + \varepsilon_i,$$

where X_i denotes the response variable, $\boldsymbol{Z}_i$ are the regressors, $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{q+1}$ are the parameter vectors and $\{\varepsilon_i\}_{i \geq 1}$ represents an innovation sequence as in Sect. 2.3.1. A least squares approach minimizing the sum of squared residuals corresponds to the vector-valued estimating function

$$\boldsymbol{H}\big((X_i, \boldsymbol{Z}_i)^T, \boldsymbol{\beta}\big) = -2\boldsymbol{Z}_i\big(X_i - \boldsymbol{Z}_i^T \boldsymbol{\beta}\big),$$

i.e., $\mathbb{X}_i = (X_i, \boldsymbol{Z}_i^T)^T$. For the mathematical results some additional assumptions on the regressors such as stationarity and independence from the innovation sequence are required.

2.3.3 Poisson autoregressive model with structural breaks

In this example we consider the Poisson autoregressive model of order one, which is a widespread model for integer-valued time series.

A Poisson autoregressive model of order one, also known as INARCH(1) model, with q change points can be described by (1), where $\{X_i^{(j)}\}$ are INARCH(1) time series with parameters $\boldsymbol{\theta}_j = (\theta_{j,1}, \theta_{j,2})^T$, $j = 1, \ldots, q+1$, i.e.,

$$X_i^{(j)} | \mathcal{F}_{i-1} \sim Poi(\lambda_i), \quad \text{with } \lambda_i = \theta_{j,1} + \theta_{j,2} X_{i-1}.$$

In this model the observation X_i conditioned on the past is Poisson distributed with a parameter λ_i depending on the past observation. A (partial) likelihood approach leads to the estimating function

$$\boldsymbol{H}((X_i, X_{i-1})^T, \boldsymbol{\theta}) = -2\boldsymbol{X}_{i-1}\left(\frac{X_i}{\boldsymbol{X}_{i-1}^T\boldsymbol{\theta}} - 1\right),$$

where $\boldsymbol{X}_{i-1} = \left(1, X_{i-1}\right)^T$, i.e., here $\mathbb{X}_i = (X_i, X_{i-1})^T$.

2.3.4 Further examples

In the context of at-most-one-change a-posteriori as well as sequential change point analysis, many procedures proposed in the literature are of this type (see Kirch and Kamgaing (2016) as well as Kirch and Kamgaing (2015) for more examples), such that their extensions to the data segmentation problem is covered by this framework.

In the context of the mean change model as in Sect. 2.3.1, in many applications the variance of the error sequence $\mathrm{Var}(\varepsilon_i) = \sigma^2$ changes in addition to the mean. Thus, it is reasonable to assume that the variance is only piecewise constant such that a change point can be caused by a mean change, a variance change or a change in both parameters. In order to detect these types of changes one can use the following vector-valued estimating function $\boldsymbol{H}(X_i, \mu, \sigma^2) = \left(X_i - \mu, (X_i - \mu)^2 - \sigma^2\right)^T$ (which relates to the method of moments but also the maximum likelihood estimator under Gaussianity assumption). A related MOSUM-Wald statistic has recently been proposed and analyzed by Messer (2022).

2.4 Threshold selection

In the following theorem, we derive the limit distribution of the maximum of the MOSUM statistics if no structural break occurs. To this end, let

$$a(x) = \sqrt{2\log(x)} \text{ and}$$
$$b(x) = 2\log(x) + \frac{p}{2}\log(\log(x)) - \log\left(\frac{2}{3}\Gamma\left(\frac{p}{2}\right)\right), \tag{7}$$

where p is the dimension of the parameter space and Γ denotes the gamma function.

Theorem 1 *Consider model (1) with no change, i.e., $q = 0$. Let Assumptions 3 on the bandwidth hold, in addition to Assumptions 1, where (b) needs to hold with $\boldsymbol{\theta} = \boldsymbol{\theta}_1$ (as defined in Assumption 2) for the MOSUM-Wald and with $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}$ for the MOSUM-score statistic. For the MOSUM-Wald statistic, let also Assumption 2 hold.*

(a) *Then, for both the MOSUM-Wald ($\ell = 1$) as well as the MOSUM-score ($\ell = 2$ with $T_{k,n}^{(2)}(G) = T_{k,n}^{(2)}(G, \widetilde{\boldsymbol{\theta}})$) statistic it holds*

$$a(n/G) \max_{k=G,\ldots,n-G} T_{k,n}^{(\ell)}(G) - b(n/G) \xrightarrow{\mathcal{D}} E$$

with $a(x)$ and $b(x)$ as in (7) and with E a Gumbel distributed random variable fulfilling $P(E \leq x) = \exp(-2\exp(-x))$.

(b) *For the MOSUM-score statistic the result remains true for data-dependent inspection parameters $\widetilde{\boldsymbol{\theta}}_n$ if Assumption 4 holds. Additionally, the assertions remain true if the long-run covariance matrix $\boldsymbol{\Sigma}_1$ is replaced by an estimator fulfilling (with $\|\cdot\|$ denoting the spectral norm of a matrix)*

$$\max_{G \leq k \leq n-G} \left\| \widehat{\boldsymbol{\Sigma}}_{k,n}^{-1/2} - \boldsymbol{\Sigma}_1^{-1/2} \right\| = o_P\left(\left(\log(n/G)\right)^{-1} \right).$$

(c) *The assertion for the MOSUM-Wald statistic remains true if both $\boldsymbol{\Sigma}_1$ and $\boldsymbol{V}_1$ are replaced by estimators $\widehat{\boldsymbol{\Sigma}}_{k,n}$, respectively, $\widehat{\boldsymbol{V}}_{k,n}$ fulfilling*

$$\max_{G \leq k \leq n-G} \left\| \boldsymbol{\Sigma}_1^{-1/2} \boldsymbol{V}(\boldsymbol{\theta}_1) - \widehat{\boldsymbol{\Sigma}}_{k,n}^{-1/2} \widehat{\boldsymbol{V}}_{k,n} \right\| = o_P\left((\log(n/G))^{-1} \right). \tag{8}$$

As threshold in our segmentation algorithm as described in Sect. 2.2 we use the asymptotic $(1-\alpha_n)$-quantiles as given in Theorem 1, i.e.,

$$D_n(\alpha_n, G) = \frac{b(n/G) + c_{\alpha_n}}{a(n/G)}, \qquad c_{\alpha_n} = -\log\log\frac{1}{\sqrt{1-\alpha_n}}, \tag{9}$$

where c_{α_n} is the $(1-\alpha_n)$-quantile of the limiting Gumbel distribution. Furthermore, we require $\alpha_n \to 0$ sufficiently slowly in the sense of:

Assumption 5 Let the sequence of significance levels α_n fulfill

$$\alpha_n \to 0 \quad \text{and} \quad c_{\alpha_n} = O(b(n/G)).$$

Remark 1 For fixed α the threshold $D_n(\alpha, G)$ can be used as a critical value in a test procedure based on MOSUM statistics for the null hypothesis of no change versus the alternative of at least one change. Proposition 1 below shows when such a test procedure has asymptotic power one. While the power is usually not as good as for the corresponding at-most-one-change statistics even in the presence of multiple changes, the latter are not as suitable for the localization of change points.

3 Consistency of the MOSUM segmentation procedures

3.1 MOSUM-Wald procedure

Part (a) of the following assumption controls the behavior of estimators that are not contaminated by change points and extends standards results from a pointwise to a uniform (within a G-environment) situation. It follows from Assumption 2 under weak assumptions on the time series and estimating functions. On the other hand, for contaminated estimators (containing observations from two segments), the estimator cannot be expected to converge to either of the best approximating parameters but will be contaminated by both (see Theorem 7 below). Part (b) of the below assumption guarantees that the estimator differs from the best approximating parameter of the jth segment if at least εG of the summands are from a neighboring regime. For sufficiently smooth estimating functions and mixing time series, the validity of the assumption will be shown in Sect. 4 below.

Assumption 6 For any $j = 1, \ldots, q$ let

$$(a) \qquad \max_{\substack{k_{j-1,n} < k \le k_{j-1,n} + G \\ k_{j,n} - 2G \le k \le k_{j,n} - G}} \sqrt{G}\, \|\hat{\theta}_{k+1,k+G} - \theta_j\| = O_P\Big(\sqrt{\log(n/G)}\Big).$$

$$(b) \qquad \sqrt{\frac{G}{\log(n/G)}} \min_{\substack{k_{j-1,n} - G \le k \le k_{j-1,n} - \varepsilon G \\ k_{j,n} - (1-\varepsilon) G \le k \le k_{j,n}}} \|\hat{\theta}_{k+1,k+G} - \theta_j\| \xrightarrow{P} \infty.$$

Because the long-run covariance matrix $\mathbf{\Gamma}_k = \boldsymbol{V}_k^{-1} \mathbf{\Sigma}_k \big(\boldsymbol{V}_k^{-1}\big)^T$ is usually unknown in applications we need to understand the influence of estimating this quantity on the procedure. This is reflected by the following assumption stating that the estimator needs to be consistent away from changes (related to (8)), while we can allow for a certain amount of misspecification close to the changes.

Assumption 7 Let $\widehat{\mathbf{\Gamma}}_{k,n}$ be an estimator sequence of $\mathbf{\Gamma}_k$ satisfying

(a) $\max_{k_{j-1,n}+G \le k \le k_{j,n}-G} \Big\| \widehat{\mathbf{\Gamma}}_{k,n}^{-1/2} - \mathbf{\Gamma}_k^{-1/2} \Big\| = o_P\big(\log(n/G)^{-1}\big)$ for any $j = 1, \ldots, q+1$,

(b) for any $j = 1, \ldots, q$ that

$$\sup_{|k-k_{j,n}|\leq(1-\varepsilon)G} \left\| \widehat{\mathbf{\Gamma}}_{k,n}^{1/2} \right\| < \infty, \qquad \sup_{|k-k_{j,n}|\leq(1-\varepsilon)G} \left\| \widehat{\mathbf{\Gamma}}_{k,n}^{-1/2} \right\| < \infty.$$

Remark 2 In Assumption 7 (a) we require a uniformly consistent covariance estimator at locations well away from change points. This assumption allows to work with the minimal threshold as discussed in Sect. 2.4 of exact order $\sqrt{\log(n/G)}$ which in turns leads to minimal assumptions on the signal strength in Assumption 6(b). In practice, it can be difficult or numerically expensive to use such estimators (see also Sects. 5 and B). In such cases, we can replace this assumption by the weaker one that

$$\max_{k_{j-1,n}+G\leq k\leq k_{j,n}-G} \left\| \widehat{\mathbf{\Gamma}}_{k,n}^{-1/2} \right\| < \infty.$$

Then, the below results remain true as long as we use a slightly larger threshold $\widetilde{D}_n(G)$ that fulfills $\widetilde{D}_n(G)/\sqrt{\log(n/G)} \to \infty$ as well as slightly strengthen Assumption 6(b) on the signal strength to

$$\frac{\sqrt{G}}{\widetilde{D}_n(G)} \min_{\substack{k_{j-1,n}-G\leq k\leq k_{j-1,n}-\varepsilon G \\ k_{j,n}-(1-\varepsilon)G\leq k\leq k_{j,n}}} \|\widehat{\boldsymbol{\theta}}_{k+1,k+G} - \boldsymbol{\theta}_j\| \xrightarrow{P} \infty.$$

We are now ready to prove a proposition, from which we will derive a consistency theorem for the MOSUM segmentation including first rates of convergence.

Proposition 1 *Let Assumptions 3 on the bandwidth hold, in addition to Assumptions 1, where (b) needs to hold with $\boldsymbol{\theta} = \boldsymbol{\theta}_j$, $j = 1, \ldots, q+1$, (as defined in Assumption 2) in addition to Assumptions 2 and 6.*

(a) *Then, for $D_n(\alpha_n, G)$ as in (9) fulfilling Assumption 5*

$$(i) \quad P\left(\max_{j=1,\ldots,q+1} \max_{k_{j-1,n}+G\leq k\leq k_{j,n}-G} T_{k,n}^{(1)}(G) < D_n(\alpha_n, G) \right) \to 1,$$

$$(ii) \quad P\left(\min_{j=1,\ldots,q} \min_{|k-k_{j,n}|\leq(1-\varepsilon)G} T_{k,n}^{(1)}(G) \geq D_n(\alpha_n, G) \right) \to 1.$$

(b) *Furthermore, the statements of (a) remain true if the long-run covariance matrix $\mathbf{\Gamma}_k$ is replaced by estimators $\widehat{\mathbf{\Gamma}}_{k,n}$ satisfying Assumption 7.*

The following theorem shows that the number of change points are estimated consistently and that the distance from each change point to the closest estimator is smaller than G.

Theorem 2 *Let the assumptions of Proposition 1 hold* (*either with known or estimated long-run covariance matrix*). *Then, as* $n \to \infty$,

$$P\left(\widehat{q}_n^{(1)} = q, \max_{1\le j\le q}\left|\widehat{k}_{j,n}^{(1)} - k_{j,n}\right| < G\right) \to 1,$$

where $\widehat{q}_n^{(1)}$ *is the estimated number of change points based on the MOSUM-Wald statistics and* $\widehat{k}_{j,n}^{(1)}$ *are the corresponding* (*ordered*) *change point estimators.*

As a special case it follows:

Remark 3 In the classical multiple change point situation, where $k_{j,n} = \lfloor \lambda_j n \rfloor$ for $0 < \lambda_1 < \ldots < \lambda_q < 1$ and $\widehat{\lambda}_{j,n}^{(1)} = \widehat{k}_{j,n}^{(1)}/n$, it holds

$$\max_{j=1,\ldots,\min(q,\widehat{q}_n^{(1)})}\left|\widehat{\lambda}_{j,n} - \lambda_{j,n}\right| = O_P\left(\frac{G}{n}\right) = o_P(1).$$

3.2 MOSUM-score procedure

The main disadvantage of the MOSUM-Wald procedure is its large computational complexity in nonlinear problems where $n - G$ estimators need to be calculated via numerical optimization. Additionally, the use of numerical procedures can cause statistical problems if distant parameters describe very similar models (corresponding to many local maxima for the numerical optimization). A prominent example of this type are neural networks.The MOSUM-score procedure on the other hand does not suffer from this problem but may (depending on the underlying model and choice of estimating function) not detect all changes if only applied with a single inspection parameter.

3.2.1 Detectability

A change in the parameter vector $\boldsymbol{\theta}$ can only be detected or localized by the MOSUM-score statistics if it causes a change in the expectation of the transformed series $\boldsymbol{H}(\mathbb{X}_i, \widetilde{\boldsymbol{\theta}})$ for the chosen inspection parameter $\widetilde{\boldsymbol{\theta}}$. Denote the corresponding set and its cardinality by

$$\begin{aligned}\tilde{Q} &= \tilde{Q}(\widetilde{\boldsymbol{\theta}}) = \left\{1 \le j \le q \,\middle|\, E\boldsymbol{H}(\mathbb{X}_1^{(j)}, \widetilde{\boldsymbol{\theta}}) \ne E\boldsymbol{H}(\mathbb{X}_1^{(j+1)}, \widetilde{\boldsymbol{\theta}})\right\},\\ \tilde{q} &= \tilde{q}(\widetilde{\boldsymbol{\theta}}) = |\tilde{Q}|.\end{aligned} \tag{10}$$

In order to formulate the below results it is helpful to relabel detectable changes $k_{j,n}$ with $j \in \tilde{Q}$ by $\tilde{k}_{j,n} = \tilde{k}_{j,n}(\widetilde{\boldsymbol{\theta}})$, $j = 1, \ldots, \tilde{q}$, where $\tilde{k}_{1,n} < \ldots < \tilde{k}_{\tilde{q},n}$.

Clearly, the number of detectable changes depends on both the choice of estimating function and the inspection parameter. Figure 2 gives an example for $\tilde{q} \ne q$,

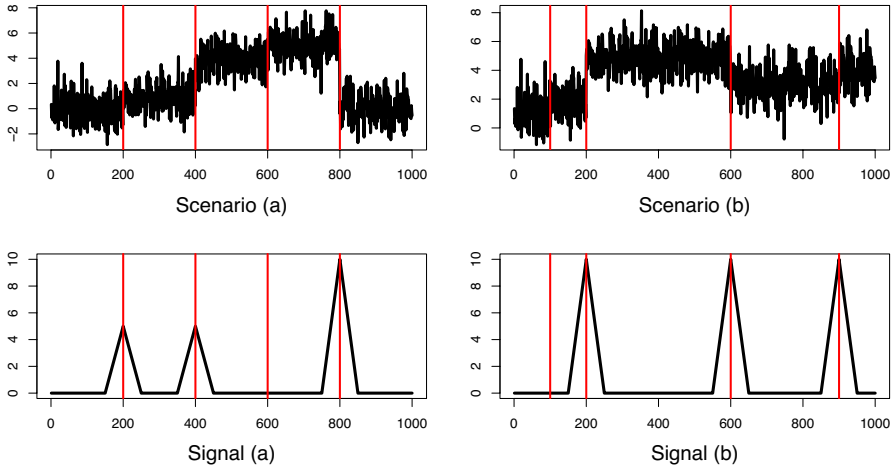


Fig. 2 The plots give two examples where $\tilde{q} \neq q$, i.e., not all changes are detectable. The upper panel shows a simulated time series, while the lower panel shows the corresponding signal of the MOSUM-score statistic with the global median as inspection parameter and the estimating function of the median. The change points are indicated by the red vertical lines

where the estimating function for the median, i.e., $H(x, \mu) = \operatorname{sgn}(x - \mu)$, was used with the global median as inspection parameter. Using a smooth strictly monotone approximation of this estimating function (compare Sect. 2.3.1) makes all changes detectable theoretically but still leads to power loss for some changes, which can be remedied by using different inspection parameters in addition to some post-processing (see Remark 4 below). This effect is illustrated by means of a simulation study as well as a data example in Sect. 5.1 below.

For the classical multiple change point situation, where $k_{j,n} = \lfloor \lambda_j n \rfloor$ for $0 = \lambda_0 < \lambda_1 < \ldots < \lambda_q < \lambda_{q+1} = 1$ a canonical data-driven inspection parameter is $\widetilde{\boldsymbol{\theta}}_n = \widehat{\boldsymbol{\theta}}_{1,n}$ defined as the solution of $\sum_{i=1}^{n} \boldsymbol{H}(\mathbb{X}_i, \boldsymbol{\theta}) \stackrel{!}{=} \mathbf{0}$.

Theorem 7 below shows that under mild conditions this data-dependent inspection parameter fulfills the required assumptions with $\widetilde{\boldsymbol{\theta}}_{0,1}$ as the unique solution of $\sum_{j=1}^{q+1} (\lambda_j - \lambda_{j-1}) E\boldsymbol{H}(\mathbb{X}_1^{(j)}, \boldsymbol{\theta}) \stackrel{!}{=} \mathbf{0}$.

Lemma 1 *Let $\boldsymbol{\theta}_j$ denote the unique zero of $E\boldsymbol{H}(\mathbb{X}_1^{(j)}, \boldsymbol{\theta})$, $j = 1, \ldots, q+1$, with $\boldsymbol{\theta}_j \neq \boldsymbol{\theta}_{j+1}$ for all $j = 1, \ldots, q$ and $q \geq 1$.*

(a) *Then, the MOSUM-score procedure with detection parameter $\widetilde{\boldsymbol{\theta}}_{0,1}$ (or a corresponding data-driven version) detects at least one change, i.e., $\widetilde{q} \geq 1$.*
(b) *If there are only two distinct regimes then all changes are detectable, i.e., $\widetilde{q} = q$.*

Remark 4 In particular, Lemma 1 shows that at least one change point is detectable if the global estimator (computed on the whole sample) is used as inspection parameter. Consequently, by recursively applying this procedure on detected segments,

all changes are detected with a much smaller computational burden than for the MOSUM-Wald procedure. An empirical illustration based on simulated data and a well-log data set is given in Sect. 5.1 below. This is of particular interest if the procedure is used for candidate generation to be combined with a pruning step as suggested by Cho and Kirch (2022). Some first considerations and results in that direction can be found in Chapter 5 of Reckrühm (2019).

3.2.2 Consistency

The MOSUM-score statistics depends on the long-run covariance matrix $\mathbf{\Sigma}_k(\widetilde{\boldsymbol{\theta}})$ as in (2) which is typically not known. We show that it can be replaced by an estimator as long as the estimator is consistent away from changes and fulfills some weaker assumptions close to a change point. The latter is important because local estimators will typically be contaminated by a close-by change point. For many models (other than mean changes) $\mathbf{\Sigma}_k(\widetilde{\boldsymbol{\theta}})$ differs from one segment to the next, such that there is is no alternative to local estimation.

Assumption 8 Let $\widehat{\mathbf{\Sigma}}_{k,n}$ satisfy

(a) $\max_{k_{j-1,n}+G\leq k\leq k_{j,n}-G} \left\| \widehat{\mathbf{\Sigma}}_{k,n}^{-1/2} - \mathbf{\Sigma}_k(\widetilde{\boldsymbol{\theta}})^{-1/2} \right\| = o_P\big(\log(n/G)^{-1}\big)$, for any $j = 1, \ldots, q+1$,

(b) For any $j = 1, \ldots, q$ it holds

$$\sup_{|k-k_{j,n}|<G} \left\| \widehat{\mathbf{\Sigma}}_{k,n}^{1/2} \right\| < \infty, \qquad \sup_{|k-k_{j,n}|<G} \left\| \widehat{\mathbf{\Sigma}}_{k,n}^{-1/2} \right\| < \infty.$$

Remark 5 As in Remark 2 it can be difficult or numerically expensive to use an estimator that is uniformly consistent away from change points. This is even more important in combination with the MOSUM-score statistics that is mainly designed to ease the computational burden of the procedure (see also Sects. 5 and B in the supplementary material). Here, too, the below results remain true as long as one uses a threshold $\widetilde{D}_n(T)$ with $\widetilde{D}_n(T)/\sqrt{\log(n/G)} \to \infty$.

In analogy to the MOSUM-Wald statistics we prove the following proposition.

Proposition 2 *Let Assumptions 3 on the bandwidth hold, in addition to Assumptions 4, thus allowing for a sequence of inspection parameters $\widetilde{\boldsymbol{\theta}}_n$. Furthermore, let Assumptions 1 hold with $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}$ (as in Assumption 4) in (b).*

(a) *Then*,

$$(i) \quad P\left(\max_{j=1,\dots,\tilde{q}+1} \max_{\tilde{k}_{j-1,n}+G\le k\le \tilde{k}_{j,n}-G} T^{(2)}_{k,n}(G) < D_n(\alpha_n, G)\right) \to 1,$$

$$(ii) \quad P\left(\min_{j=1,\dots,\tilde{q}} \min_{|k-\tilde{k}_{j,n}|\le(1-\varepsilon)\,G} T^{(2)}_{k,n}(G) \ge D_n(\alpha_n, G)\right) \to 1.$$

(b) *Furthermore, the statements of* (*a*) *remain true if the long-run covariance matrix* $\boldsymbol{\Sigma}_k(\widetilde{\boldsymbol{\theta}})$ *is replaced by estimators* $\widehat{\boldsymbol{\Sigma}}_{k,n}$ *satisfying Assumption* 8.

From this we conclude in the following theorem that the number of change points are estimated consistently and that the distance from each change point to the closest estimator is smaller than G.

Theorem 3 *Let the assumptions of Proposition* 2 *hold* (*either with known or estimated long-run covariance matrix*). *Then, as* $n \to \infty$,

$$P\left(\widehat{q}^{(2)}_n(\widetilde{\boldsymbol{\theta}}_n) = \tilde{q}(\widetilde{\boldsymbol{\theta}}), \max_{1\le j\le \tilde{q}} \left|\widehat{k}^{(2)}_{j,n}(\widetilde{\boldsymbol{\theta}}_n) - \tilde{k}_{j,n}(\widetilde{\boldsymbol{\theta}})\right| < G\right) \to 1,$$

where $\widehat{q}^{(2)}_n(\widetilde{\boldsymbol{\theta}}_n)$ *is the estimated number of change points based on the MOSUM-score statistics with inspection parameters* $\widetilde{\boldsymbol{\theta}}_n$ *and* $\widehat{k}^{(2)}_{j,n}(\widetilde{\boldsymbol{\theta}}_n)$ *are the corresponding* (*ordered*) *change point estimators.*

Remark 6 In the classical multiple change point situation, where $\tilde{k}_{j,n} = \lfloor \tilde{\lambda}_j\, n \rfloor$ for $0 < \tilde{\lambda}_1 < \ldots < \tilde{\lambda}_{\tilde{q}} < 1$ and $\widehat{\lambda}^{(2)}_{j,n} = \widehat{k}^{(2)}_{j,n}/n$, it holds

$$\max_{j=1,\dots,\min(\tilde{q},\widehat{q}^{(2)}_n)} |\widehat{\lambda}_{j,n} - \tilde{\lambda}_{j,n}| = O_P\left(\frac{G}{n}\right) = o_P(1).$$

3.2.3 Localization rates

Localization rates quantify the distance between true and estimated change points. For a bounded number of change points the minimax optimal localization rates are known to be constant, i.e., it holds $\max_j |\widehat{k}_{j,n} - k_j| = O_P(1)$; see, e.g., Lemma 2 in Wang et al. (2020). In this section we show that the MOSUM-score estimators match these minimax optimal localization rates.

This is important because the MOSUM-score procedure is particularly well suited as a candidate generating algorithm to be combined with a pruning step (see Cho and Kirch (2022) for a corresponding discussion in the context of the mean change model). This is due to its computational efficiency and the fact that combined with several inspection parameters all changes can be detected. For such pruning methodology to work it is essential that the candidate generating algorithm produces estimators that are sufficiently close to the true change points.

In this section, we work with $\widehat{k}^{(2)}_{j,n}(\widetilde{\boldsymbol{\theta}}_n; \widehat{\boldsymbol{\Psi}}_{j,n})$ as in (6), i.e., for $j \in \widetilde{Q}$

$$\widehat{k}_{j,n}^{(2)}(\widetilde{\boldsymbol{\theta}}_n;\widehat{\boldsymbol{\Psi}}_{j,n}) = \underset{v_{j,n}\le k\le w_{j,n}}{\arg\max}\, \boldsymbol{M}_{\widetilde{\boldsymbol{\theta}}_n}(k)^T \widehat{\boldsymbol{\Psi}}_{j,n}^{-1} \boldsymbol{M}_{\widetilde{\boldsymbol{\theta}}_n}(k).$$

We need the following assumptions on the matrices $\widehat{\boldsymbol{\Psi}}_{j,n}$.

Assumption 9 $\widehat{\boldsymbol{\Psi}}_{j,n}$ is a sequence (possibly depending on the data) of symmetric positive definite matrices with

$$\|\widehat{\boldsymbol{\Psi}}_{j,n}^{-1}\| = O_P(1), \qquad \|\widehat{\boldsymbol{\Psi}}_{j,n}^{1/2}\| = O_P(1).$$

Effectively, the assumption guarantees that the use of $\widehat{\boldsymbol{\Psi}}_{j,n}$ does not kill the signal nor does it explode the noise.

Furthermore, we need to slightly strengthen the previous assumptions by additionally requiring:

Assumption 10

(a) With $\widetilde{\boldsymbol{\theta}}$ as in Assumption 4 we require that for any $\xi_n \to \infty$ and any $j = 1, \dots, \tilde{q}_n + 1$, it holds

(*i*) for $m = \tilde{k}_{j-1,n}(\widetilde{\boldsymbol{\theta}}) + G,\ \tilde{k}_{j,n}(\widetilde{\boldsymbol{\theta}}) - G,\ \tilde{k}_{j,n}(\widetilde{\boldsymbol{\theta}})$ that

$$\max_{\xi_n\le k\le G} \frac{1}{k} \left\| \sum_{i=m-k+1}^{m} \left(\boldsymbol{H}\left(\mathbb{X}_i^{(j)}, \widetilde{\boldsymbol{\theta}}_n\right) - \boldsymbol{H}\left(\mathbb{X}_i^{(j)}, \widetilde{\boldsymbol{\theta}}\right)\right)\right\| = o_P(1),$$

(*ii*) for $m = \tilde{k}_{j-1,n}(\widetilde{\boldsymbol{\theta}}),\ \tilde{k}_{j-1,n}(\widetilde{\boldsymbol{\theta}}) + G,\ \tilde{k}_{j,n}(\widetilde{\boldsymbol{\theta}}) - G$ that

$$\max_{\xi_n\le k\le G} \frac{1}{k} \left\| \sum_{i=m+1}^{m+k} \left(\boldsymbol{H}\left(\mathbb{X}_i^{(j)}, \widetilde{\boldsymbol{\theta}}_n\right) - \boldsymbol{H}\left(\mathbb{X}_i^{(j)}, \widetilde{\boldsymbol{\theta}}\right)\right)\right\| = o_P(1).$$

(b) The following backward law of large numbers holds for any j and any sequence $\xi_n \to \infty$

$$\max_{\xi_n\le k\le G} \frac{1}{k} \left\| \sum_{i=G-k+1}^{G} \left(\boldsymbol{H}\left(\mathbb{X}_i^{(j)}, \widetilde{\boldsymbol{\theta}}\right) - E\boldsymbol{H}\left(\mathbb{X}_i^{(j)}, \widetilde{\boldsymbol{\theta}}\right)\right)\right\| = o_P(1).$$

Remark 7 The forward version (as in (a)(ii)) is missing in (b) because it follows directly from a strong law of large number which (after changing the probability space) follows directly from the invariance principle as in Assumption 1 (b). However, the backward version does not follow from that assumption alone - unless the data is i.i.d. Nevertheless, most sequences will allow for strong ergodic theorems even in backward direction. For example, mixing conditions as in Regularity conditions 1(b) are symmetric in the sense that the backward sequence is also mixing with the same rates. Consequently, both a forward and backward invariance principle holds and (b) follows from that.

In the following theorem we derive the minimax optimal localization rates improving upon Theorem 3.

Theorem 4 *Let the assumptions of Theorem 3 hold in addition to Assumptions 9 – 10. Then, it holds*

$$\max_{1\leq j\leq \min(\tilde{q},\widehat{q}_n^{(2)}(\widetilde{\boldsymbol{\theta}}_n))} \left|\widehat{k}_{j,n}^{(2)}(\widetilde{\boldsymbol{\theta}}_n;\widehat{\boldsymbol{\Psi}}_{j,n}) - \tilde{k}_{j,n}(\widetilde{\boldsymbol{\theta}})\right| = O_P(1).$$

3.3 Outlook: possible extensions beyond single-scale MOSUM

Single-scale MOSUM procedures can only deal with homogeneous change points, where all changes are of the same signal strength. On the other hand, in practice, one encounters often heterogeneous situations, where there are both frequent changes with larger change magnitudes (as, e.g., measured by a suitable difference in the parameter values) as well as changes with smaller change magnitudes surrounded by longer stretches of stationarity. For a mathematical definition in the context of changes in the mean we refer to Cho and Kirch (2022), Definition 2.1. While the first kind can typically be detected with smaller bandwidths, longer bandwidths (i.e., longer effective sample sizes for change detection) are required for the latter. Using several bandwidths leads to duplicate and spurious change point estimators from which final candidates need to be selected. For the classical sample mean-based procedures, several authors have discussed such post-processing steps (for a detailed discussion thereof, we refer to Cho and Kirch (2021), Section 3.2.4):

Fryzlewicz (2014) proposes a randomized local procedure (wild binary segmentation (WBS)) for candidate generation along with suitable post-processing methods. Subsequently, variations thereof, including new post-processing techniques has been proposed such as Baranowski et al. (2019), Fryzlewicz (2020), Fryzlewicz (2023). Cho and Kirch (2022) propose and mathematically analyze a two-stage procedure which includes a candidate generating step in addition to a post-processing step based on the BIC-information criterion: In particular, they prove consistency of the post-processing procedure in combination with different candidate generating methods including mean-based MOSUM for different bandwidths under weaker assumptions than is possible for the single-scale MOSUM. Figure 3 in Eichinger and Kirch (2018) gives empirical evidence that for data sets with heterogeneous change points a single bandwidth is not sufficient to achieve consistency.

In combination with such a post-processing step, one can take full advantage of the MOSUM-score procedure as a computationally fast candidate generating method as the same inspection parameter can be used with a variety of bandwidths (see also Remark 4). Here, several data-driven inspection parameters can be used limiting the required number of numerical optimization procedures. The MOSUM-Wald procedure on the other hand requires a separate (numerical) computation of parameter estimators for each bandwidth and each point in

time making the MOSUM-Wald procedure computationally potentially highly problematic.

However, it is not straightforward how to adapt the BIC-based post-processing to this general framework as BIC is effectively motivated from a likelihood approach in the context of normal data with multiple changes in the mean but constant variance across time. Because, here, we are in a general framework allowing for robust methods via corresponding estimating functions as well as complicated time series structures, a detailed discussion or analysis of such post-processing procedures in this framework is left for future work. Some first considerations in that direction can be found in Chapter 5 of Reckrühm (2019).

4 Regularity conditions for sufficiently smooth estimating functions

The results of the previous sections were obtained under certain high-level assumptions, that need to be verified separately for each underlying time series structure as well as estimating function. On the other hand, many estimating functions are either sufficiently smooth or can be approximated to any degree of accuracy by sufficiently smooth functions. Therefore, we will verify these high-level assumptions exemplarily for sufficiently smooth estimating function for strongly mixing time series under moment conditions. This includes the i.i.d. situation as a special case.

The invariance principles of Assumption 1 (b) follow from the following regularity conditions:

Regularity Condition 1 For $j = 1, \ldots, q+1$

(a) let there exist a $\tilde{\nu} > 0$ such that $0 < E\left\|\boldsymbol{H}(\mathbb{X}_1^{(j)}, \boldsymbol{\theta})\right\|^{2+\tilde{\nu}} < \infty$ for the $\boldsymbol{\theta}$ of interest,

(b) let $\{\mathbb{X}_t^{(j)}\}$ be stationary and strongly mixing sequences of random vectors with a strong mixing coefficient $\alpha(\cdot)$ satisfying $\alpha(n) = O(n^{-\beta})$ for some $\beta > 1 + 2/\tilde{\nu}$ as $n \to \infty$, where $\tilde{\nu}$ is as (a).

Theorem 5

(a) *Under Regularity Conditions 1 Assumption 1 hold for the same $\boldsymbol{\theta}$ and some $\nu > 0$ depending on β, $\tilde{\nu}$ and the dimension.*

(b) *Additionally, Assumption 10 (b) holds if Regularity Conditions 1 are fulfilled with $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}$.*

In the following we will always assume that the bandwidth G fulfills Assumption 3 for the above suitable ν.

In order to prove the necessary assumptions for the consistency results of the above MOSUM statistics the following additional regularity conditions on the estimating function are sufficient.

Regularity Condition 2

(a) Let $\boldsymbol{\Theta} \subset \mathbb{R}^p$ be a compact parameter space and $\boldsymbol{H} = (H_1, \ldots, H_p)^T$ twice differentiable such that for some $\tilde{\nu} > 0$

$$(i) \quad E\left\|\nabla \boldsymbol{H}(\mathbb{X}_1^{(j)}, \boldsymbol{\theta})\right\|^{2+\tilde{\nu}} < \infty \qquad \text{for all } \boldsymbol{\theta} \in \boldsymbol{\Theta},$$

$$(ii) \quad E \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\|\nabla^2 H_l(\mathbb{X}_1^{(j)}, \boldsymbol{\theta})\right\|^{2+\tilde{\nu}} < \infty, \quad \text{for all } l = 1, \ldots, p,$$

for all $j = 1, \ldots, q+1$.

(b) Let $\boldsymbol{V}_{(j)}(\boldsymbol{\theta}) = E\Big(\nabla \boldsymbol{H}(\mathbb{X}_1^{(j)}, \boldsymbol{\theta})\Big)^T$ be a regular matrix for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ fulfilling for all $j = 1, \ldots, q+1$

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\|\boldsymbol{V}_{(j)}(\boldsymbol{\theta})^{-1}\right\| < \infty.$$

(c) Let $E \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\|\nabla H_l(\mathbb{X}_1^{(j)}, \boldsymbol{\theta})\right\| < \infty$, $j = 1, \ldots, q+1$.

For specific examples the above assumptions such as the compactness assumption can be weakened. The important special case of linear regression models has been discussed in detail in Reckrühm (2019), Chapter 3.2, under standard assumptions for linear regression.

Theorem 6 *Let Regularity Condition 1 hold with $\boldsymbol{\theta} = \boldsymbol{\theta}_j$ as defined in Assumption 2 identifiably unique, $j = 1, \ldots, q+1$, in addition to Regularity Condition 2 (a) and (b).*

(a) *Then Assumptions 2 and Assumption 6 (a) hold.*
(b) *If additionally Regularity Condition 2 (c) hold, we get Assumption 6 (b).*

The MOSUM-score procedure requires an inspection parameter $\widetilde{\boldsymbol{\theta}}$ which can also be data-dependent $\widetilde{\boldsymbol{\theta}}_n$. We will show the validity of the assumptions for situations where $\widetilde{\boldsymbol{\theta}}_n$ is a $\sqrt{n}$-consistent estimator for some $\widetilde{\boldsymbol{\theta}}$ in Theorem 8.

Here, we discuss the choice $\widehat{\boldsymbol{\theta}}_{a,b}$ defined as the unique zero of $\sum_{i=a}^{b} \boldsymbol{H}(\mathbb{X}_i, \boldsymbol{\theta}) \stackrel{!}{=} \mathbf{0}$, which has advantageous properties for detectability (compare Sect. 3.2.1) in particular if combined with a subsequent pruning step.

Theorem 7 *Let the classical multiple change point situation with $k_{j,n} = \lfloor \lambda_j n \rfloor$ for $0 = \lambda_0 < \lambda_1 < \ldots < \lambda_q < \lambda_{q+1} = 1$ hold. Consider $a = \lfloor \gamma_a n \rfloor$ and $b = \lfloor \gamma_b n \rfloor$ for some $0 \le \gamma_a < \gamma_b \le 1$. Define $\widetilde{\boldsymbol{\theta}}_{\gamma_a,\gamma_b}$ as the unique solution of $\sum_{j=1}^{q+1} (\min(\lambda_j, \gamma_b) - \max(\lambda_{j-1}, \gamma_a))_+ E\boldsymbol{H}(\mathbb{X}_1^{(j)}, \boldsymbol{\theta}) \stackrel{!}{=} \mathbf{0}$, where $x_+ = \max(x, 0)$. Furthermore, let Regularity Conditions 1 hold with $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}_{\gamma_a,\gamma_b}$ as well as 2 (b) and (c). Then,*

Table 1 Percentage of simulations with a change point estimator in the given intervals ('detection rate') based on the MOSUM-score statistic with the median-like estimating function

G	[80, 120]	[180, 220]	[580, 620]	[880, 920]
Inspection Parameter: Sample median $\hat{\mu}_{1,1000}$				
20	0.019	1.000	0.935	0.142
50	0.343	1.000	1.000	0.665
Inspection Parameter: Sample median $\hat{\mu}_{1,200}$				
20	0.158	0.985	0.380	0.026
50	0.659	1.000	0.999	0.404

$$\sqrt{n}\left(\hat{\theta}_{a,b} - \widetilde{\theta}_{\gamma_a,\gamma_b}\right) = O_P(1).$$

Finally, we show the validity of the necessary assumptions for the MOSUM-score procedure under suitable regularity conditions on the estimating function.

Theorem 8 *Let $\widetilde{\theta}_n$ and $\widetilde{\theta}$ with $\sqrt{n}(\widetilde{\theta}_n - \widetilde{\theta}) = O_P(1)$ in addition to Regularity Condition 1 with $\theta = \widetilde{\theta}$ hold. Additionally, let Regularity Condition 2 (a) and (c) hold. Then Assumption 4 as well as 10 (a) hold.*

5 Empirical comparison

For the mean change situation based on the sample mean as in Sect. 2.3.1 extensive simulation studies can be found in Eichinger and Kirch (2018); Cho and Kirch (2022) and Meier et al. (2021). In this section, we want to give a proof of concept beyond the mean by including a variety of examples in the simulation study, where we focus on aspects that differ from the mean situation. In this section we aim at sheding some additional light on the detectability issue discussed in Sect. 3.2.1 using the example of a mean change in combination with the median-like estimator. In Sections A and B (in the supplementary material) we focus on differences between the MOSUM-Wald and MOSUM-score statistics in the empirical performance as well as their computation time. In particular, this includes the example of linear regression in Section A, where the estimator is known analytically, as well as the example of Poisson autogression in Section B, where this is not the case and all estimators need to be obtained numerically.

All simulations are based on 1000 repetitions unless otherwise stated, ε as in (5) is set to 0.2 and the threshold from Sect. 2.4 with $\alpha_n = 5\%$ is used.

5.1 Median-like estimator and detectability

In Sect. 3.2.1 the detectability issue of the MOSUM-score procedure was discussed in detail. Here, we shed some additional light onto the problem by a corresponding empirical study and the analysis of well-log data where we use the median-like

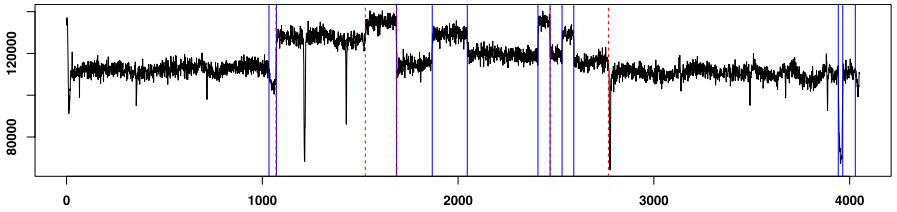
estimating function as discussed in Sect. 2.3.1 with the sample median (based on different stretches of the original data) as inspection parameter. If instead the median-like estimator is used as an inspection parameter, the detection rates as given in Table 1 only differ in the third digit. However, using the sample median as implemented in `R` instead of the median-like estimator (implemented using the `uniroot`-function from the `R`-package `stats`) saved around 1/3 of computation time in our otherwise identical simulation.

The procedure is implemented using the `mosum` function from the R-package `mosum` applied to the transformed data sequence $H(X_t, \widehat{\theta}_n)$. In particular, the variance is estimated empirically with the mosum-window estimator that is the default in the package, see Meier et al. (2021). Thus, the signal is effectively given by the difference in the expectation of the transformed data $EH(X_t, \widetilde{\theta})$ with the noise corresponding to the distribution of $H(X_t, \widetilde{\theta})$ where $\widetilde{\theta}$ is the limit of $\widehat{\theta}_n$.

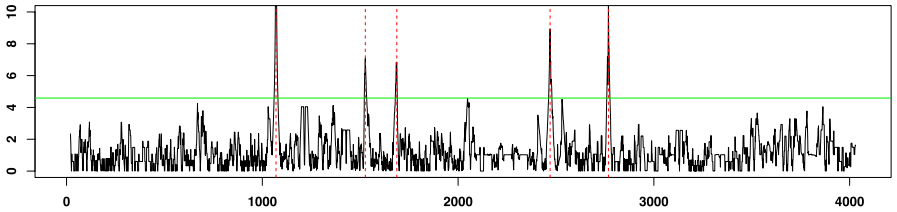
To elaborate we consider a noise sequence of 1000 independent standard normally distributed random variables which are then centered around a mean sequence of 1, 2, 5, 3, 4 with change points at 100, 200, 600 and 900. Consequently, the global median will be somewhere between 3 and 4. As such the second (jumping from mean 2 to mean 5) and to a slightly lesser degree the third change point (jumping from 5 to 3) should be well visible to our statistic applied with the global median, while the last one (jumping from 3 to 4) should be less visible. This is also confirmed by the empirical detection rates as reported in Table 1. The first change point should be almost invisible as it jumps from 1 to 2 with only few observations exceeding the global median. While the detection rates in Table 1 are indeed lower than for the other change points, they are well above the 5% threshold with detection rates of approximately 20% (for $G = 20$) and even of approximately 35% (with $G = 50$). The reasons are twofold: First of all, more observations from the second regime can be expected to cross the threshold than for the first one. Secondly, with the median-like estimation function, the data transformation is strictly monotonic such that any mean difference remains theoretically detectable with any inspection parameter. Furthermore, the transformation that dampens the signal (if both means are on the same side of the inspection parameter) will also dampen the noise (i.e., variance) of the signal. As soon as the sample median of the first 200 observations is used as an inspection parameter, which will be between 1 and 2, the detection rates for the first change point increases while the ones for the last change point decreases illustrating the discussion in Sect. 3.3 and Remark 4.

5.2 Analysis of well-log data

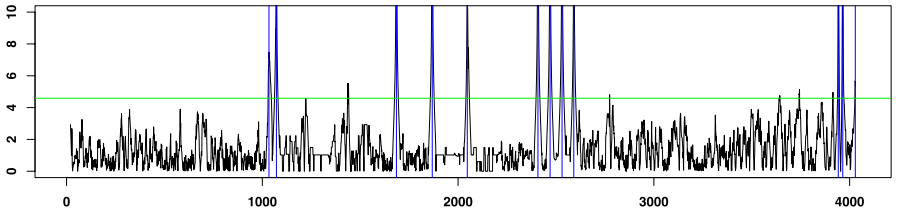
One of the advantages of using the median in combination with the median-like estimating function is the fact that they are robust with respect to outliers. Robust data segmentation methodology has recently obtained some interest, for example, Knoblauch et al. (2018), Fearnhead and Rigaill (2019) and Li and Yu (2021) present robust change point procedures and illustrate their usefulness using well-log data (see Fig. 3a): This



(a) Well-log data with detected change points



(b) MOSUM-Score statistic with global sample median as inspection parameter



(c) MOSUM-Score statistic with median between 1070 and 2767 as inspection parameter

Fig. 3 Well-log data with segmentation obtained from the median-like MOSUM-Score statistic with different inspection parameters. The vertical dashed red lines give the detected change points using the global median as inspection parameter, while the vertical solid blue lines give the detected change points using as an inspection parameter the median of the stretch starting at 1070 and ending at 2767

geophysical data[1] is collected from a probe being lowered into a bore-hole, where change points occur when the probe moves from one rock strata to another. The data notably contains several outliers, which have been removed before the change point analysis in earlier works such as (Ruanaidh and Fitzgerald (1996), Section 5.7.2) where the data was first discussed, or Fearnhead and Clifford (2003), Fearnhead (2006), Adams and MacKay (2007), Wyse et al. (2011), Ruggieri and Antonellis (2016).

Using the median-like estimating function in combination with the sample median yields a robust procedure such that we can use this methodology on the original data set including the outliers. We use a bandwidth of $G = 20$ and first apply the procedure using the global median as inspection parameter (see Fig. 3b). This results in the detection of 5 change points leaving some more undetected (as predicted by the discussion in Sect. 3.2.1). In a second step, we repeat the procedure using now the sample median between the first (at 1070) and last (at 2768) detected

[1] Available at https://github.com/alan-turing-institute/rbocpdms/.

change point as inspection parameter (see Fig. 3c). This results in the detection of 9 additional change points, which together result in a very reasonable segmentation of the data (see Fig. 3a). Out of the three change points detected by both inspection parameters in one case the estimator was exactly at the same location (at 2470), while the other two were shifted by 2 time points (one in either direction). This further emphasizes the need for suitable post-processing methodology that can deal with this type of duplicate estimators obtained from different inspection parameters or different bandwidths (see also Remark 4). Using additionally inspection parameters obtained from the data up to the first, respectively, starting at the last change point (detected with the global median at 1070, respectively, 2768) does not yield any additional change point estimators but some more duplicated ones (8 that are exactly equal, and 6 slightly shifted).

Concerning the last three detected change points in Fig. 3a and c, the question may arise as to whether these are truly change points or in fact outliers. Indeed, the stretch between 3942 and 3965 as been removed by some of the previous works as outliers but not all of them. From an asymptotic perspective the difference between serial outliers and multiple change points is the following: For serial outliers the length of the outlier stretch is bounded while the jump size increases to infinity (sufficiently fast), whereas in change point analyses the jump size is bounded while the distance to the next neighboring change point increases. Indeed, typically (see, e.g., Cho and Kirch (2022), Definition 2.1) one needs that this distance times the squared jump size diverges, where one can even allow the jump size to go to zero. In a finite sample situation, the distinction is less clear, explaining why some authors have considered this stretch of 23 very small observations as two change points, while others have considered them serial outliers.

The MOSUM procedure is robust with respect to classical outliers that can occur in both directions in an isolated way. Longer (with respect to the bandwidth G) stretches of serial outliers are in some sense indistinguishable from segments of length of order G and thus may be detected by the procedure: While the breakdown point of the median is at 50%, smaller stretches of outliers (in the same direction) of length γG, $0 < \gamma < 0.5$, will lead to the situation that the sample median of the window of G observations (including the γG serial outliers) will estimate the $(0.5 \pm 0.5\,\gamma/(1-\gamma))$-quantile instead which typically will differ from the median of the neighboring stretch that does not contain any serial outliers. This also explains the last change point in Fig. 3a and c because there the 'outlier' stretch (if one decides to classify it as such) has a length of around 10 and we use a bandwidth of only 20. On the other hand, a suitable choice of bandwidth G in combination with the choice of ε in (5) will make the procedure robust with respect to sufficiently short outlier stretches. The latter effect can, e.g., be seen in Fig. 3c, where some of the other outlier stretches lead to significant points in the MOSUM-score statistic that are sufficiently isolated to not be considered change points by the procedure due to the choice of $\varepsilon = 0.2$ in (5).

With larger bandwidths the long stretches at the beginning and the end of the sequence are also segmented to some extend, which is consistent with observations made by some previous authors (even after outlier removal) because these stretches

seem not to be mean-stationary which will cause location segmentation procedures to fit a step-function.

6 Conclusion

In this work, we propose and analyze a data segmentation procedure in a unified framework including likelihood-based as well as robust methodology for complex nonlinear time series such as Poisson regression or neural network-based (auto-) regression. To account for the diversity of models included in the framework, we work with high-level assumptions in Sects. 2 and 3, that need to be verified on a case-by-case basis in general. For sufficiently smooth estimating functions, this high-level assumptions are shown to be valid under standard moment conditions in Sect. 4. Assumptions 7 for the MOSUM-Wald as well as 8 for the MOSUM-score procedures and their relaxations in Remarks 2 and 5 on the covariance estimators are very general. Consequently, the choice of covariance estimators has only a minimal influence on the asymptotic properties of the corresponding change point estimators, however, it is critical for the small sample behavior: On the one hand (away from change points), the choice of the threshold in combination with the covariance estimator greatly influences the rate of spuriously detected change points. On the other hand (close to change points), this combination determines the detectability rate for a given signal strength. Estimators for the covariance depend crucially on the given model and estimating function, such that they need to be chosen separately for every situation.

Another contribution of this paper is the analysis of two different procedures: The MOSUM-Wald methodology yields better results with a single bandwidth but at the cost of a higher computation time in particular in combination with nonlinear models requiring numerical optimization methodology. Furthermore, the MOSUM-Wald procedure is prone to errors in a situation where many local maxima are present in the optimization procedure such as is well known for neural network approximations. The reason is that two parameters that are far away in the parameter space can describe almost the same model but will be mistaken for different models by the MOSUM-Wald procedure. The MOSUM-score procedure on the other hand is robust to a multiple mode situations and in addition is computationally much lighter even for nonlinear models. On the other hand, with a single inspection parameter it might not be able to detect all change points present (see Sect. 3.2.1 as well as 5.1), thus, future work is needed to solve this problem (see Sect. 3.3).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10463-023-00892-4. It contains additional simulations as well as the detailed proofs. Furthermore, R-code to generate Figs. 1, 2 and 3 is also provided.

Acknowledgements This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 314838170, GRK 2297 MathCoRe. Claudia Kirch would also like to thank the Isaac Newton Institute for Mathematical Sciences for support and hospitality during the program 'Statistical scalability' when work on this paper was undertaken.

References

Adams, R. P., MacKay, D. J. (2007). Bayesian online changepoint detection. arXiv preprint arXiv:0710.3742.

Aggarwal, R., Inclan, C., Leal, R. (1999). Volatility in emerging stock markets. *Journal of Financial and Quantitative Analysis, 34*(1), 33–55.

Aue, A., Horváth, L. (2013). Structural breaks in time series. *Journal of Time Series Analysis, 34*(1), 1–16.

Baranowski, R., Chen, Y., Fryzlewicz, P. (2019). Narrowest-over-threshold detection of multiple change-points and change-point-like features. *Journal of the Royal Statistical Society, Series B, 81*, 649–672.

Bauer, P., Hackl, P. (1980). An extension of the mosum technique for quality control. *Technometrics, 22*(1), 1–7.

Braun, J. V., Braun, R. K., Müller, H.-G. (2000). Multiple changepoint fitting via quasilikelihood, with application to dna sequence segmentation. *Biometrika, 87*(2), 301–314.

Chen, J., Gupta, A. K. (2012). *Parametric statistical change point analysis: With applications to genetics, medicine, and finance* (2nd ed.). Boston: Springer Science & Business Media.

Cho, H., Kirch, C. (2021). Data segmentation algorithms: Univariate mean change and beyond. *Econometrics and Statistics*. https://doi.org/10.1016/j.ecosta.2021.10.008

Cho, H., Kirch, C. (2022). Two-stage data segmentation permitting multi-scale changepoints, heavy tails and dependence. *Annals of the Institute of Statistical Mathematics, 74*, 653–684.

Chu, C.-S.J., Hornik, K., Kaun, C.-M. (1995). Mosum tests for parameter constancy. *Biometrika 82*(3), 603–617.

Csörgö, M., Horváth, L. (1997). *Limit theorems in change-point analysis*. Hoboken: John Wiley & Sons Inc.

Eichinger, B., Kirch, C. (2018). A mosum procedure for the estimation of multiple random change points. *Bernoulli, 24*(1), 526–564.

Fearnhead, P. (2006). Exact and efficient bayesian inference for multiple changepoint problems. *Statistics and Computing, 16*(2), 203–213.

Fearnhead, P., Clifford, P. (2003). On-line inference for hidden markov models via particle filters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65*(4), 887–899.

Fearnhead, P., Rigaill, G. (2019). Changepoint detection in the presence of outliers. *Journal of the American Statistical Association, 114*(525), 169–183.

Fearnhead, P., Rigaill, G. (2020). Relating and comparing methods for detecting changes in mean. Stat, e291.

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics, 42*(6), 2243–2281.

Fryzlewicz, P. (2020). Detecting possibly frequent change-points: Wild Binary Segmentation 2 and steepest-drop model selection. *Journal of the Korean Statistical Society, 49*, 1–44.

Fryzlewicz, P. (2023). Narrowest significance pursuit: inference for multiple change-points in linear models. *Journal of the American Statistical Association*, 1–14.

He, X., Shao, Q.-M. (1996). A general bahadur representation of m-estimators and its application to linear regression with nonstochastic designs. *The Annals of Statistics, 24*(6), 2608–2630.

Horváth, L., Rice, G. (2014). Extensions of some classical methods in change point analysis. *Test, 23*(2), 219–255.

Hušková, M. (1990). Asymptotics for robust mosum. *Commentationes Mathematicae Universitatis Carolinae, 31*(2), 345–356.

Hušková, M. (2013). Robust change point analysis. *Robustness and complex data structures* (pp. 171–190). Berlin: Springer.

Hušková, M., Slabỳ, A. (2001). Permutation tests for multiple changes. *Kybernetika, 37*(5), 605–622.

Killick, R., Eckley, I. A., Ewans, K., Jonathan, P. (2010). Detection of changes in variance of oceanographic time-series using changepoint analysis. *Ocean Engineering, 37*, 1120–1126.

Killick, R., Fearnhead, P., Eckley, I. A. (2012). Optimal detection of change-points with a linear computational cost. *Journal of the American Statistical Association, 107*(500), 1590–1598.

Kirch, C., Kamgaing, J. T. (2012). Testing for parameter stability in nonlinear autoregressive models. *Journal of Time Series Analysis, 33*(3), 365–385.

Kirch, C., Kamgaing, J. T. (2015). On the use of estimating functions in monitoring time series for change points. *Journal of Statistical Planning and Inference, 161*, 25–49.

Kirch, C., Kamgaing, J. T. (2016). Detection of change points in discrete valued time series. In R. A. Davis, S. H. Holan, R. Lund, N. Ravishanker (Eds.), *Handbook of discrete valued time series* (pp. 219–244). Boca Raton: CRC Press.

Kirch, C., Klein, P. (2023). Moving sum data segmentation for stochastics processes based on invariance. *Statistica Sinica, 33*, 1–20. https://doi.org/10.5705/ss.202021.0048

Kirch, C., Weber, S. (2018). Modified sequential change point procedures based on estimating functions. *Electronic Journal of Statistics, 12*(1), 1579–1613.

Knoblauch, J., Jewson, J. E., Damoulas, T. (2018). Doubly robust bayesian inference for non-stationary streaming data with β-divergences. *Advances in Neural Information Processing Systems*, 31 .

Korkas, K. K., Fryzlewicz, P. (2017). Multiple change-point detection for non-stationary time series using wild binary segmentation. *Statistica Sinica, 27*(1), 287–311.

Li, M., Yu, Y. (2021). Adversarially robust change point detection. *Advances in Neural Information Processing Systems, 34*, 22955–22967.

Meier, A., Cho, H., Kirch, C. (2021). mosum: A package for moving sums in change point analysis. *Journal of Statistical Software, 97*, 1–42.

Messer, M. (2022). Bivariate change point detection: joint detection of changes in expectation and variance. *Scandinavian Journal of Statistics, 49*(2), 886–916.

Messer, M., Kirchner, M., Schiemann, J., Roeper, J., Neininger, R., Schneider, G. (2014). A multiple filter test for the detection of rate changes in renewal processes with varying variance. *The Annals of Applied Statistics, 8*(4), 2027–2067.

Mohammad-Djafari, A., Féron, O. (2006). Bayesian approach to change points detection in time series. *International Journal of Imaging Systems and Technology, 16*(5), 215–221.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika, 41*, 100–115.

Reckrühm, K. (2019). Estimating multiple structural breaks in time series: A generalized mosum approach based on estimating functions (Doctoral dissertation, Otto-von-Guericke–Universität Magdeburg, Faculty of Mathematics). Retrieved from https://doi.org/10.25673/13832

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

Ruggieri, E., Antonellis, M. (2016). An exact approach to bayesian sequential change point detection. *Computational Statistics & Data Analysis, 97*, 71–86.

Tartakovsky, A. (2019). *Sequential change detection and hypothesis testing: General non-iid stochastic models and asymptotically optimal rules*. New York: CRC Press.

Tartakovsky, A., Nikiforov, I., Basseville, M. (2014). *Sequential analysis: Hypothesis testing and change-point detection*. New York: CRC Press.

Truong, C., Oudre, L., Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing, 167*, 107299.

Vostrikova, L. J. (1981). Detecting disorder in multidimensional random processes. *Soviet Mathematics Doklady, 24*, 55–59.

Wang, D., Yu, Y., Rinaldo, A. (2020). Univariate mean change point detection: Penalization, cusum and optimality. *Electronic Journal of Statistics, 14*(1), 1917–1961.

Wyse, J., Friel, N., Rue, H. (2011). Approximate simulation-free bayesian inference for multiple change-point models with dependence within segments. *Bayesian Analysis, 6*(4), 501–528.

Yau, C. Y., Zhao, Z. (2016). Inference for multiple change points in time series via likelihood ratio scan statistics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 78*(4), 895–916.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.