Multiscale change point detection for dependent data

Holger Dette, Theresa Schüler
Fakultät für Mathematik
Ruhr-Universität Bochum
44799 Bochum, Germany

Mathias Vetter
Mathematisches Seminar
Christian-Albrechts-Universität zu Kiel
24098 Kiel, Germany

November 14, 2018

Abstract

In this paper we study the theoretical properties of the simultaneous multiscale change point estimator (SMUCE) proposed by Frick et al. (2014) in regression models with dependent error processes. Empirical studies show that in this case the change point estimate is inconsistent, but it is not known if alternatives suggested in the literature for correlated data are consistent. We propose a modification of SMUCE scaling the basic statistic by the long run variance of the error process, which is estimated by a difference-type variance estimator calculated from local means from different blocks. For this modification we prove model consistency for physical dependent error processes and illustrate the finite sample performance by means of a simulation study.

Keywords and phrases: Change point detection, multiscale methods, physical dependent processes
AMS Subject Classification: 62M10, 62G08,

1 Introduction

The problem of detecting multiple abrupt changes in the structural properties of a time series and to split the data into several "stationary" segments has been of interest to statisticians for many decades. An efficient a posteriori change-point detection rule enables the researcher to analyze data under the assumption of piecewise-stationarity and has numerous applications including bioinformatics, neuroscience, genetics, the analysis of speech signals, financial, and climate data. Because of its importance the literature on the subject is very vast and we refer exemplarily to the work of Yao (1988), Bai and Perron (1998, 2003), Braun et al. (2000), Lavielle and Moulines

(2000), Kolaczyk and Nowak (2005), Davis et al. (2006), Harchaoui and Lévy-Leduc (2010), Ciuperca (2011, 2014), Killick et al. (2012), Fryzlewicz (2014), Matteson and James (2014), Cho and Fryzlewicz (2015), Preuss et al. (2015), Yau and Zhao (2016), Haynes et al. (2017), Korkas and Fryzlewicz (2017) and Chakar et al. (2017). This list of references is by no means complete and further references can be found in the cited literature.

The focus of the present paper is on the simultaneous multiscale change point estimator (SMUCE), which was introduced recently in a seminal paper of Frick et al. (2014) to identify multiple changes in the mean structure of the sequence

$$Y_i = \vartheta^*\left(\tfrac{i}{n}\right) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where $\vartheta^* : [0,1] \to \mathbb{R}$ is a piecewise constant function and $\varepsilon_1, \ldots, \varepsilon_n$ are independent identically distributed centered Gaussian random variables. Note that these authors considered distributions from a one-parametric exponential family with a piecewise constant parameter ϑ^*, but for the sake of brevity we restrict ourselves to the location scale model, which corresponds to the Gaussian case. The SMUCE procedure controls the probability of overestimating the true number of change points, and it is also possible to give bounds for the probability of underestimation. Moreover, one can construct asymptotic honest confidence sets for the unknown step function ϑ^* and its change points. The method has turned out to be very successful and has therefore been extended in various directions. For example, Pein et al. (2017b) consider model (1.1) with a heteroscedastic Gaussian noise process. Li et al. (2016) argue that in situations with low signal to noise ratio or with many change-points compared to the number of observations SMUCE necessarily leads to a conservative estimate and propose to control the false discovery instead of the family wise error rate. More recently Li et al. (2018) extend the procedure to certain function classes beyond step functions in a nonparametric regression setting.

The present paper is devoted to the analysis of SMUCE in the location scale model (1.1) with a piecewise constant regression function under more general assumptions on the error process. We are particularly interested in the situation where the errors are neither Gaussian nor independent. If the sample size is reasonably large and the errors are independent, SMUCE is relatively robust because it is based on local means which are asymptotically Gaussian due to the CLT. However, the independence of the errors is more crucial and ignoring this assumption may lead to serious errors in the estimation procedure. This is illustrated in Figure 1, where we display a typical estimate of the signal (upper left panel) by the modification of SMUCE proposed in Tecuapetla-Gómez and Munk (2017) for m-dependent errors (lower left panel). The data generating process is an ARMA$(2,6)$ process. We observe that the modification still produces a function with too many jumps. The lower right panel shows the estimate proposed in this paper, which seems to work better. The upper right panel shows the performance of SMUCE, which clearly overestimates the true number of change points. A more detailed comparison will be presented in Section 4.

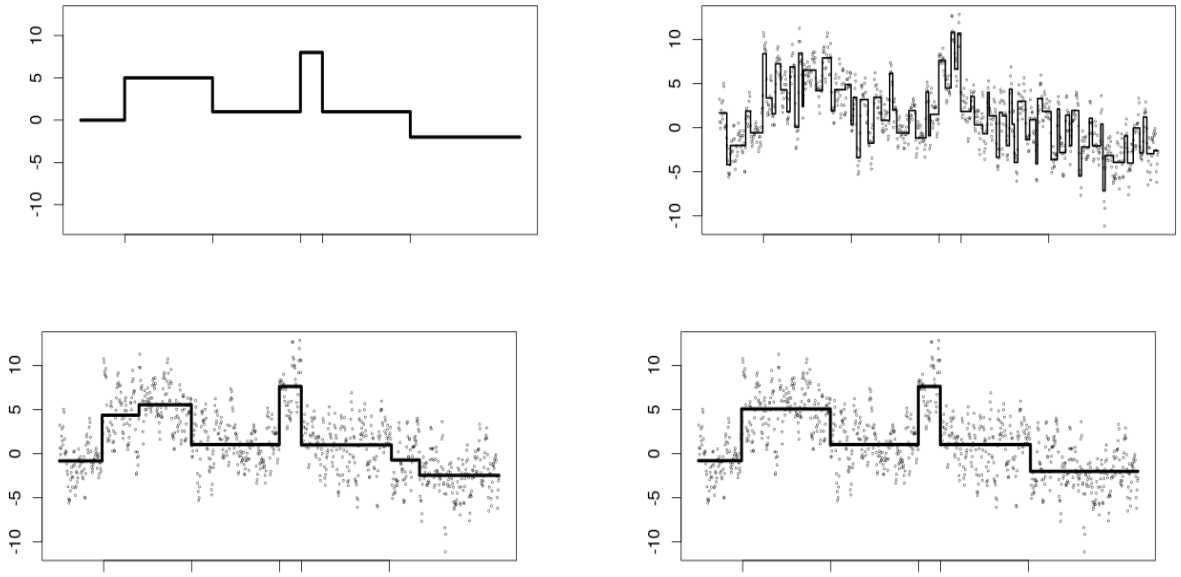

Figure 1: *Different estimates of a piecewise constant signal in model* (1.1) *with an* $ARMA(2,6)$ *error process. Upper left panel: true function. Upper right panel: SMUCE. Lower left panel: estimate proposed in Tecuapetla-Gómez and Munk (2017). Lower right panel: estimate proposed in this paper.*

The reason for the differences consists in the fact that in the case of dependent data all described procedures require a reliable estimate of the long run variance of the error distribution. Tecuapetla-Gómez and Munk (2017) demonstrate by means of a simulation study that the problem can easily be addressed for m-dependent errors using difference based estimators [see Hall et al. (1990) or Dette et al. (1998)]. Their approach provides a solution for a specific error structure and we see the improvement in Figure 1. However, from a practical point of view the method requires a good choice of m, and the example indicates that this procedure might not work well for other dependence structures. More importantly, from a theoretical point of view rigorous statements regarding the performance of SMUCE in models with more general (stationary) error processes are missing. It turns out that results of this type are substantially more difficult to obtain and are–to our best knowledge–not available in the literature so far.

In this paper we address this problem and prove consistency of SMUCE with an appropriately modified variance estimator under the assumption that the error process $\{\varepsilon_i\}_{i\in\mathbb{Z}}$ is a physical system in the sense of Wu (2005). This includes such important examples as ARMA or GARCH processes. We also avoid any distributional assumptions regarding the errors ε_i except the existence of moments. In Section 2 we introduce the model and the modification of the SMUCE procedure to address general time dependent error processes. Roughly speaking, we have to

define consistent estimates of the long run variance

$$\sigma_\star^2 := \sum_{k\in\mathbb{Z}} \mathrm{Cov}(\varepsilon_0, \varepsilon_k) \ , \tag{1.2}$$

which address the fact that the regression function may be only piecewise constant and not constant. This is achieved by a two step estimator which is defined as a difference based estimator of local averages. The asymptotic properties of the modified procedure are established in Section 3. We prove that the number of change points is identified with probability converging to 1 and that all change points are estimated consistently. The finite sample properties are investigated in Section 4 by means of a simulation study. Finally, all proofs and technical details are deferred to an appendix.

2 Multiscale change point detection for dependent data

We begin with a brief review of the simultaneous multiscale change point estimator (SMUCE) as introduced by Frick et al. (2014), where we directly address the problem of dependent data. Throughout this paper let

$$\vartheta^*(t) := \sum_{k=0}^{K^*} \theta_k^* \mathbb{1}_{[\tau_k^*, \tau_{k+1}^*)}(t) \tag{2.1}$$

denote the "true" unknown signal in model (1.1), where K^* is the (unknown) number of change points, $0 = \tau_0^* < \tau_1^* < \ldots < \tau_{K^*}^* < \tau_{K^*+1}^* = 1$ are the change point locations, and $\theta_0^*, \ldots, \theta_{K^*}^*$ are the function values of ϑ^*. We summarize the change point locations in a vector

$$J(\vartheta^*) = (\tau_1^*, \ldots, \tau_{K^*}^*)$$

of dimension $|J(\vartheta^*)|$. For the sake of simplicity we restrict ourselves to estimators of the form $\hat\vartheta(\cdot) = \sum_{k=0}^{\hat K} \hat\theta_k \mathbb{1}_{[\hat\tau_k, \hat\tau_{k+1})}(\cdot)$ where the estimates $\hat\tau_k$ of the change point locations only attain values at the sampling points $0, \frac{1}{n}, \ldots \frac{n-1}{n}, 1$ and denote the set of these functions by $\mathcal{S}_n$. Following Frick et al. (2014) we propose to test for a candidate step function $\vartheta(\cdot) = \sum_{k=0}^{K} \theta_k \mathbb{1}_{[\tau_k, \tau_{k+1})}(\cdot) \in \mathcal{S}_n$ on each interval $[i/n, j/n]$ where ϑ is constant whether ϑ^* is constant on this interval as well with the same value as ϑ. For this purpose we use the multiscale statistic

$$V_n(Y, \vartheta) = \max_{0\le k\le K} \max_{\substack{n\tau_k \le i \le j < n\tau_{k+1} \\ j-i+1 \ge nc_n}} \left\{ \frac{1}{\hat\sigma_\star} \sqrt{j-i+1} \left| \overline{Y}_i^j - \theta_k \right| - \sqrt{2 \log \frac{en}{j-i+1}} \right\}, \tag{2.2}$$

where $\{c_n\}_{n\in\mathbb{N}}$ is a positive sequence converging to 0,

$$\overline{Y}_i^j := \frac{1}{j-i+1} \sum_{\ell=i}^{j} Y_\ell$$

is a local mean and $\hat{\sigma}_\star^2$ is an appropriate estimator of the long run variance (1.2), which will be defined later. The estimator of the piecewise constant function ϑ^* is then required to minimize the number of change points over the acceptance region of this multiscale test. More precisely, for a fixed threshold q chosen according to the (asymptotic) null distribution of V_n the step function estimator $\hat{\vartheta}$ is required to fulfil a data fit claim of the form

$$V_n(Y, \hat{\vartheta}) \leq q \ ,$$

and to satisfy simultaneously a parsimony requirement concerning its number of change points. This is achieved by first estimating the number of change points K^* by

$$\hat{K} = \hat{K}(V_n, q) = \inf_{\substack{\vartheta \in \mathcal{S}_n \\ V_n(Y,\vartheta) \leq q}} |J(\vartheta)|.$$

Next, we identify among all suitable candidate step functions the one which provides the best fit to the data, that is

$$\hat{\vartheta} = \operatorname*{argmin}_{\vartheta \in \mathcal{C}(V_n, q)} \sum_{i=1}^{n} \Big(Y_i - \vartheta\big(\tfrac{i}{n}\big) \Big)^2 \ , \tag{2.3}$$

where

$$\mathcal{C}(V_n, q) := \{\vartheta \in \mathcal{S}_n \ : \ |J(\vartheta)| = \hat{K} \text{ and } V_n(Y, \vartheta) \leq q\}$$

is a "confidence set" of all functions in $\mathcal{S}_n$ satsifying the multiscale criterion with a minimal number of change points. The estimator can be efficiently computed by a dynamic program and is implemented with the function *stepFit* in the R-package *stepR* [see Pein et al. (2017a)].

The appropriate estimation of the long run variance $\sigma_\star^2$ is crucial for a good performance of SMUCE if it is applied to correlated data, and for this purpose we propose a two step procedure as considered in Wu and Zhao (2007). We divide the sample in $m_n = \lfloor \frac{n}{k_n} \rfloor$ blocks $\{Y_1, \ldots, Y_{k_n}\}$, $\{Y_{k_n+1}, \ldots, Y_{2k_n}\}, \ldots, \{Y_{(m_n-1)k_n+1}, \ldots, Y_{m_n k_n}\}$ of length k_n and calculate local averages

$$A_i := \frac{1}{k_n} \sum_{j=1}^{k_n} Y_{j+ik_n},$$

to mimic the dependence structure of the data. Secondly, we use the difference based estimate

$$\hat{\sigma}_\star^2 := \frac{k_n}{2(m_n - 1)} \sum_{i=1}^{m_n - 1} |A_i - A_{i-1}|^2 \,, \tag{2.4}$$

to eliminate the signal. Here k_n increases with the sample size in order to achieve the correct asymptotic behaviour. For details see Proposition 3.1 below, where we prove the consistency of this estimate.

In the Gaussian case the only difference to the SMUCE procedure regards the use of the long run variance estimator. Note, however, that we will discuss arbitrary dependent error processes, not necessarily Gaussian, in which case the asymptotic analysis of the procedure is substantially more difficult. This analysis will be carefully carried out in the following Section 3. The finite sample properties of the new multiscale method are investigated by means of a simulation study in Section 4.

3 Asymptotic properties

Consider the location scale model (1.1) with a stationary error process $\varepsilon = \{\varepsilon_i\}_{i\in\mathbb{Z}}$ such that $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}[\varepsilon_i] = \sigma^2 > 0$. For the asymptotic analysis of the multiscale procedure introduced in Section 2 we assume that ε is a physical system as introduced in Wu (2005). This means that there exists a sequence of independent identically distributed random variables $\{\eta_i\}_{i\in\mathbb{Z}}$ with values in some measure space $\mathcal{S}$ and a measurable function $G : \mathcal{S}^{\mathbb{N}} \to \mathbb{R}$ such that for all $i \in \mathbb{Z}$

$$\varepsilon_i = G(\ldots, \eta_{i-1}, \eta_i) \ .$$

As pointed out by Wu (2011), physical systems include many of the commonly used time series models such as ARMA and GARCH processes.

In the following discussion let $p \geq 1$ and define for a random variable X (in the case of its existence) $||X||_p = \left(\mathbb{E}[|X|^p]\right)^{1/p}$. If $\|\varepsilon_i\|_p < \infty$ we consider the physical dependence measure

$$\delta_{i,p} := ||\varepsilon_i - \varepsilon_i^\star||_p,$$

where the random variable $\varepsilon_i^\star$ is defined by $\varepsilon_i^\star = G(\ldots, \eta_{-1}, \eta_0', \eta_1, \ldots, \eta_i)$ and η_0' is an independent copy of η_0. We also define the quantity

$$\Delta_{m,p} := \sum_{i=m}^{\infty} \delta_{i,p}, \quad m = 1, 2, \ldots$$

and call a system $\{\varepsilon_i\}_{i\in\mathbb{Z}}$ p-strong stable if $\Delta_{0,p} < \infty$ [see Wu (2005)]. It can be shown that for a 2-strong stable process $\{\varepsilon_i\}_{i\in\mathbb{Z}}$ the covariance function is absolutely summable and thus the long run variance in (1.2) exists [see e.g. Wu and Phoumaradi (2009)]. A further quantity that we will make use of is the so-called projection operator, which for $i \in \mathbb{Z}$ is given by

$$P_i \cdot := \mathbb{E}[\cdot \,|\, \mathcal{F}_i] - \mathbb{E}[\cdot \,|\, \mathcal{F}_{i-1}],$$

where $\mathcal{F}_i = (\ldots, \eta_{i-1}, \eta_i)$. It is shown in Wu (2011) that for a 2-strong stable process $\{\varepsilon_i\}_{i\in\mathbb{Z}}$ the long run variance (1.2) can be represented as $\sigma_\star^2 = \mathbb{E}[(\sum_{j=0}^{\infty} P_0\varepsilon_j)^2]$.

For the statement of the asymptotic properties in this section we will make the following basic assumptions

(A1) $\|\varepsilon_i\|_4 < \infty$

(A2) $\Delta_{0,4} < \infty$ and $\sum_{i=1}^{\infty} i\delta_{i,2} < \infty$

(A3) $\Delta_{m,3} = \mathcal{O}(m^{-\gamma})$ for some $\gamma > 0$

Assumption (A3) is used to construct a simultaneous Gaussian approximation of the partial sums of the errors ε_i (see Section 5 for details). Assumption (A2) is needed for a proof of the first result of this section, which establishes the consistency of the estimator (2.4) for the long run variance with an explicit rate. For its precise statement we introduce the notation $a_n \asymp b_n$ for two sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$, which means that

$$0 < \liminf_{n\to\infty} |a_n/b_n| \leq \limsup_{n\to\infty} |a_n/b_n| < \infty.$$

Proposition 3.1 *Consider the nonparametric regression model* (1.1) *with a piecewise constant regression function* (2.1). *If assumptions (A1) and (A2) are satisfied and* $k_n \asymp n^{1/3}$, *we have for the estimator in* (2.4)

$$\hat{\sigma}_\star - \sigma_\star = \mathcal{O}_{\mathbb{P}}\left(n^{-1/3}\right),$$

where $\sigma_\star^2$ *is the long run variance in* (1.2).

Throughout this paper we will always assume that $k_n \asymp n^{1/3}$, if the long run variance estimator (2.4) is used. Our first main result shows that the asymptotic null distribution of the statistic V_n does not change in the case of dependent observations.

Theorem 3.2 *Consider the nonparametric regression model* (1.1) *with piecewise constant regression function* (2.1). *If assumptions (A1)–(A3) are satisfied with* $\gamma > 1/2$ *in (A3),* $c_n \to 0$ *and*

$$\lim_{n\to\infty} \frac{(\log n)^3}{n^{m(\gamma)} c_n} = 0\ , \tag{3.1}$$

where $m(\gamma) = \frac{2\gamma-1}{1+6\gamma}$, *then it holds*

$$V_n(Y, \vartheta^*) \xrightarrow{\mathcal{D}} \max_{0\leq k\leq K^*} \sup_{\tau_k^* \leq s < t \leq \tau_{k+1}^*} \left\{ \frac{|B(t) - B(s)|}{\sqrt{t-s}} - \sqrt{2\log\frac{e}{t-s}} \right\} \text{ as } n \to \infty,$$

where $\{B(t)\}_{t\in[0,1]}$ *denotes a standard Brownian motion.*

With exactly the same arguments as given in Frick et al. (2014), we can assure for given $\alpha \in (0,1)$ that

$$\lim_{n\to\infty} \mathbb{P}\left(\hat{K}(V_n, q) > K^*\right) \leq \alpha, \tag{3.2}$$

where q is chosen as the $(1-\alpha)$–quantile of

$$M := \sup_{0\leq s\leq t\leq 1} \left\{ \frac{|B(t) - B(s)|}{\sqrt{t-s}} - \sqrt{2\log\frac{e}{t-s}} \right\}. \tag{3.3}$$

Note that the distribution of M coincides with the asymptotic distribution in Theorem 3.2, if the function ϑ^* is constant (that is $K^* = 0$). We also obtain from Theorem 3.2 and the definition of $\hat{K}$ that the probability of overestimating the number of change points becomes arbitrarily small with an increasing sample size.

Corollary 3.3 *If the assumptions from Theorem 3.2 are satisfied and $q_n \to \infty$, we have*

$$\lim_{n\to\infty} \mathbb{P}\left(\hat{K}(V_n, q_n) > K^*\right) = 0.$$

The following result shows that the probability of underestimating the true number of change points also converges to 0 for an increasing sample size.

Theorem 3.4 *If the assumptions from Theorem 3.2 hold and the sequence $\{q_n\}_{n\in\mathbb{N}}$ fulfils*

$$q_n = \mathrm{o}\left(\sqrt{n}\right) \tag{3.4}$$

as $n \to \infty$, then it follows that

$$\lim_{n\to\infty} \mathbb{P}\left(\hat{K}(V_n, q_n) < K^*\right) = 0.$$

Combining Corollary 3.3 and Theorem 3.4 yields model selection consistency.

Corollary 3.5 *If the same assumptions as in Theorem 3.4 are satisfied and $q_n \to \infty$, then it follows that*

$$\lim_{n\to\infty} \mathbb{P}\left(\hat{K}\left(V_n, q_n\right) = K^*\right) = 1.$$

Under appropriate assumptions, the change point locations of ϑ^* are estimated correctly. More precisely, we have the following result.

Theorem 3.6 *If the assumptions from Theorem 3.2 hold and the sequence* $\{q_n\}_{n\in\mathbb{N}}$ *additionally fulfils* $q_n \to \infty$ *and*

$$q_n + \sqrt{2\log\frac{e}{c_n}} = \mathrm{o}\left(\sqrt{nc_n}\right) \tag{3.5}$$

as $n \to \infty$*, it follows that*

$$\lim_{n\to\infty} \mathbb{P}\Big(\sup_{\vartheta\in\mathcal{C}(V_n,q_n)} \max_{\tau^*\in J(\vartheta^*)} \min_{\tau\in J(\vartheta)} |\tau^* - \tau| > c_n \Big) = 0.$$

In particular we have for $k = 1, \ldots, K^*$

$$\lim_{n\to\infty} \mathbb{P}\Big(\sup_{\vartheta\in\mathcal{C}(V_n,q_n)} |\tau_k^* - \tau_k| > c_n \Big) = 0.$$

4 Finite sample properties

In this section we compare the finite sample performance of the change point estimator developed and analyzed in Section 3 with SMUCE and the change point estimator proposed by Tecuapetla-Gómez and Munk (2017) for m-dependent errors. These authors use the abbreviation JUSD for their procedure and we will use the notation DepSMUCE for the procedure (2.3) developed in this paper. The sample size is $n = 1000$ and all results are based on 1000 simulation runs. For DepSMUCE, we consider a block length of $k = 10$. Concerning the change point estimator JUSD, it is necessary to specify a value for m. The R-package *dbacf* [see Tecuapetla-Gómez (2015)] provides a graphical procedure to choose m which is used throughout the simulation study.

We compare the deviations between the estimated and the true number of change points, and the mean deviation of $|K^* - \hat{K}|$. Concerning the data fit, we compute the mean squared error

$$\mathrm{MSE}(\hat{\vartheta}) := \frac{1}{n}\sum_{i=1}^{n} \left(\vartheta^*\left(\tfrac{i}{n}\right) - \hat{\vartheta}\left(\tfrac{i}{n}\right)\right)^2$$

and mean absolute deviation

$$\mathrm{MAE}(\hat{\vartheta}) := \frac{1}{n}\sum_{i=1}^{n} \left|\vartheta^*\left(\tfrac{i}{n}\right) - \hat{\vartheta}\left(\tfrac{i}{n}\right)\right|,$$

respectively. Furthermore, we also present histograms of the estimated locations of the changes for all three estimators.

All procedures depend sensitively on the threshold q in the definition of the change point estimator and we investigate three different choices of q. More precisely, considering (3.2), we choose

the significance level α as 0.1, 0.5 and 0.9 and set q as the $(1-\alpha)$–quantile of the distribution of the random variable M in (3.3). Since this quantile cannot be derived directly, we perform Monte Carlo simulations of the test statistic $V_n(Y, \vartheta^*)$ with $\vartheta^* \equiv 0$ and independent standard normal distributed errors, i.e. $\varepsilon_i \sim \mathcal{N}(0,1)$ (based on 10000 repetitions). This is exactly the same procedure as in the R-package *stepR* [see Pein et al. (2017a)].

First, we illustrate that SMUCE is relatively robust to weak dependencies but it does not yield satisfactory results when the innovations exhibit a stronger dependence. To this end, we consider two MA(1) error processes with different MA parameters. Let

$$\varepsilon_i = \eta_i + \kappa\eta_{i-1}, \quad i \in \mathbb{Z}, \tag{4.1}$$

where $\{\eta_i\}_{i\in\mathbb{Z}}$ is a sequence of standard normal distributed errors. We consider the cases $\kappa = 0.1$ and $\kappa = 0.3$, respectively, and assume that the function ϑ^* in model (1.1) has $K^* = 5$ change points at locations

$$(\tau_1^*, \tau_2^*, \tau_3^*, \tau_4^*, \tau_5^*) = (101/1000, 301/1000, 501/1000, 551/1000, 751/1000). \tag{4.2}$$

The corresponding function intensities are given by

$$(\theta_0^*, \theta_1^*, \theta_2^*, \theta_3^*, \theta_4^*, \theta_5^*) = (0, 1, 0, 2, 0, -1). \tag{4.3}$$

$\hat{K} - K^*$	≤ -3	-2	-1	0	$+1$	$+2$	$\geq +3$
SMUCE(0.1)	0.000	0.000	0.001	0.980	0.019	0.000	0.000
SMUCE(0.5)	0.000	0.000	0.000	0.760	0.209	0.031	0.000
SMUCE(0.9)	0.000	0.000	0.000	0.238	0.343	0.267	0.152
DepSMUCE(0.1)	0.000	0.000	0.117	0.883	0.000	0.000	0.000
DepSMUCE(0.5)	0.000	0.000	0.009	0.988	0.003	0.000	0.000
DepSMUCE(0.9)	0.000	0.000	0.000	0.946	0.053	0.001	0.000
JUSD(0.1)	0.075	0.065	0.125	0.702	0.027	0.004	0.003
JUSD(0.5)	0.020	0.024	0.080	0.745	0.087	0.020	0.023
JUSD(0.9)	0.003	0.004	0.033	0.631	0.168	0.085	0.076

Table 1: *Proportion of estimated numbers of change points (the true number of change points is $K^* = 5$) in model* (1.1) *with step function defined by* (4.2) *and* (4.3) *and an MA(1) error process defined in* (4.1) *with* $\kappa = 0.1$.

Tables 1 and 2 show the performance of SMUCE, DepSMUCE and JUSD in the estimation of the number of change points for different values of α. For example, in Table 1 we display results for model (4.1) with $\kappa = 0.1$ and we observe that DepSMUCE estimates the correct number of change points in 98.8% of the cases if we work with $\alpha = 0.5$. Considering the first three rows of Table 1, it can be seen that SMUCE performs relatively well if $\alpha = 0.1$. However for $\alpha = 0.5$ DepSMUCE already shows some improvement and for $\alpha = 0.9$ DepSMUCE and JUSD

$\hat{K} - K^*$	≤ -3	-2	-1	0	$+1$	$+2$	$\geq +3$
SMUCE(0.1)	0.000	0.000	0.001	0.619	0.302	0.066	0.012
SMUCE(0.5)	0.000	0.000	0.000	0.069	0.184	0.262	0.486
SMUCE(0.9)	0.000	0.000	0.000	0.000	0.010	0.025	0.965
DepSMUCE(0.1)	0.001	0.043	0.356	0.600	0.000	0.000	0.000
DepSMUCE(0.5)	0.000	0.000	0.048	0.947	0.005	0.000	0.000
DepSMUCE(0.9)	0.000	0.000	0.007	0.919	0.070	0.004	0.000
JUSD(0.1)	0.231	0.124	0.179	0.446	0.018	0.001	0.001
JUSD(0.5)	0.072	0.077	0.167	0.618	0.043	0.016	0.007
JUSD(0.9)	0.010	0.016	0.111	0.615	0.156	0.060	0.032

Table 2: *Proportion of estimated numbers of change points (the true number of change points is $K^* = 5$) in model* (1.1) *with step function defined by* (4.2) *and* (4.3) *and an MA(1) error process defined in* (4.1) *with* $\kappa = 0.3$.

show a better performance because they are constructed to address dependency in the data. The advantages of these two procedures become even more visible in Table 2, where we consider a stronger dependence, that is $\kappa = 0.3$. In this case SMUCE tends to overestimate the true number of change points. We also observe a better performance of DepSMUCE compared to JUSD, which often estimates a too small number of change points. Our findings are confirmed by Table 3, where we display the average MSE, MAE, and the average value of $|K^* - \hat{K}|$. Figure 2 shows histograms of the estimated change point locations for $\alpha = 0.5$. The comparatively bad performance of JUSD can be explained by the fact that it requires the specification of the order of the MA(m) process. We observed that the data driven procedure to choose m from the R-package *dbacf* [see Tecuapetla-Gómez (2015)] does not work well for small MA parameters. A simulation study with $\kappa = 0.5$, which is not included here for the sake of brevity, shows that JUSD works much better for larger MA parameters, and as a consequence the behaviour of DepSMUCE and JUSD becomes more similar.

	$\kappa = 0.1$			$\kappa = 0.3$		
	$\lvert K^* - \hat{K}\rvert$	MSE	MAE	$\lvert K^* - \hat{K}\rvert$	MSE	MAE
SMUCE(0.1)	0.020	0.018	0.060	0.475	0.033	0.093
SMUCE(0.5)	0.271	0.019	0.064	2.569	0.045	0.118
SMUCE(0.9)	1.407	0.024	0.077	6.488	0.063	0.145
DepSMUCE(0.1)	0.117	0.025	0.072	0.446	0.064	0.139
DepSMUCE(0.5)	0.012	0.018	0.060	0.053	0.031	0.088
DepSMUCE(0.9)	0.056	0.018	0.060	0.084	0.030	0.085
JUSD(0.1)	0.549	0.050	0.109	1.226	0.117	0.209
JUSD(0.5)	0.397	0.031	0.082	0.647	0.069	0.145
JUSD(0.9)	0.705	0.024	0.073	0.577	0.044	0.109

Table 3: *Average of $|K^* - \hat{K}|$, MSE, and MAE of different estimates in the MA*(1)*-model* (4.1).

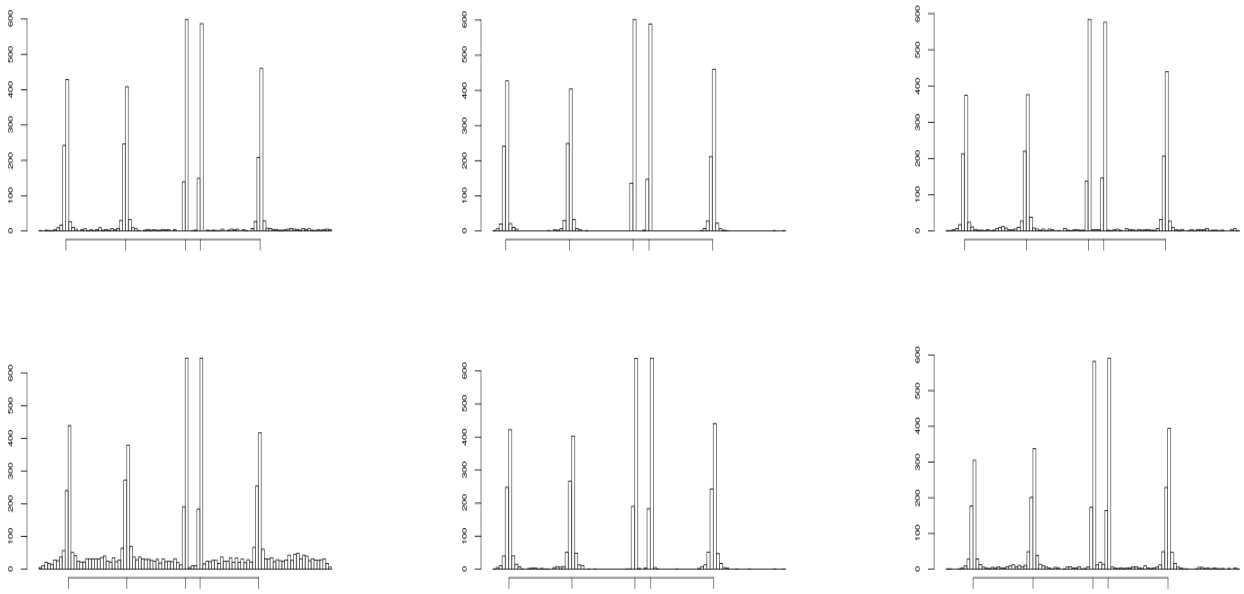

Figure 2: *Histograms of estimated change point locations for different estimators. First row: MA*(1) *error process with* $\kappa = 0.1$. *Second row: MA*(1) *error process with* $\kappa = 0.3$. *Left column: SMUCE. Middle column: DepSMUCE. Right column: JUSD. The "true" change points are located at* 101*,* 301*,* 501*,* 551*, and* 751.

This observation is also confirmed in our next example, where we consider an MA(4) error process with relatively large parameters, that is

$$\varepsilon_i = \eta_i + 0.9\eta_{i-1} + 0.8\eta_{i-2} + 0.7\eta_{i-3} + 0.6\eta_{i-4}, \quad i \in \mathbb{Z}\ . \tag{4.4}$$

Here $\{\eta_i\}_{i\in\mathbb{Z}}$ denotes again a sequence of independent standard normal distributed errors. We assume that the function ϑ^* in model (1.1) has $K^* = 5$ change points at the locations given in (4.2) and that the corresponding function intensities are given by

$$(\theta_0^*, \theta_1^*, \theta_2^*, \theta_3^*, \theta_4^*, \theta_5^*) = (0, 3, 0, 4, 0, -3). \tag{4.5}$$

The data driven rule in the R-package *dbacf* works well and determines $m = 4$ for JUSD correctly in all of the iterations. Table 4 shows the performance of SMUCE, DepSMUCE, and JUSD. For example, if $\alpha = 0.5$, DepSMUCE estimates the number K^* of change points correctly in 80.6% of the cases, while it underestimates K^* by 1 in 17.3% of the cases. The first three rows show that SMUCE is not able to correctly estimate the number of change points in the case of an MA(4)

$\hat{K} - K^*$	≤ -3	-2	-1	0	$+1$	$+2$	$\geq +3$
SMUCE(0.1)	0.000	0.000	0.000	0.000	0.000	0.000	1.000
SMUCE(0.5)	0.000	0.000	0.000	0.000	0.000	0.000	1.000
SMUCE(0.9)	0.000	0.000	0.000	0.000	0.000	0.000	1.000
DepSMUCE(0.1)	0.020	0.138	0.511	0.330	0.001	0.000	0.000
DepSMUCE(0.5)	0.000	0.006	0.173	0.806	0.015	0.000	0.000
DepSMUCE(0.9)	0.000	0.000	0.024	0.856	0.114	0.006	0.000
JUSD(0.1)	0.025	0.188	0.511	0.275	0.001	0.000	0.000
JUSD(0.5)	0.000	0.006	0.145	0.812	0.037	0.000	0.000
JUSD(0.9)	0.000	0.000	0.013	0.709	0.240	0.032	0.006

Table 4: *Proportion of estimated numbers of change points (the true number of change points is $K^* = 5$) in model* (1.1) *with step function defined by* (4.2) *and* (4.5) *and an MA(4) error process defined in* (4.4).

	MA(4)			ARMA(2,6)		
	$\lvert K^* - \hat{K}\rvert$	MSE	MAE	$\lvert K^* - \hat{K}\rvert$	MSE	MAE
SMUCE(0.1)	43.845	1.787	1.016	59.950	4.174	1.592
SMUCE(0.5)	56.842	2.041	1.108	73.800	4.550	1.684
SMUCE(0.9)	67.865	2.208	1.166	85.582	4.798	1.743
DepSMUCE(0.1)	0.848	0.861	0.584	0.516	1.534	0.778
DepSMUCE(0.5)	0.200	0.418	0.364	0.064	0.646	0.465
DepSMUCE(0.9)	0.150	0.319	0.322	0.115	0.586	0.449
JUSD(0.1)	0.963	0.929	0.619	1.422	1.720	0.879
JUSD(0.5)	0.194	0.401	0.357	2.181	1.042	0.637
JUSD(0.9)	0.336	0.343	0.341	3.979	1.019	0.630

Table 5: *Average of $|K^* - \hat{K}|$, MSE, and MAE of different estimates under the same model assumptions as in Table 4 and Table 6.*

error process. Of course, SMUCE is designed for independent data, but it always estimates a much larger number of change points than 5. In contrast, JUSD and DepSMUCE perform substantially better if they are used with $\alpha = 0.5$. In particular, they yield very similar results and DepSMUCE is able to compete with JUSD, which is specially designed for m-dependent processes (note that we used the correct m in the simulations). Similar observations can be made for the estimation error (see the left part of Table 5). These observations are also supported by the upper part of Figure 3 which shows the histograms of the estimated change point locations. DepSMUCE and JUSD are able to identify the locations correctly in most of the cases and show a rather similar behaviour. On the other hand, SMUCE is not reliable for the estimation of the signal in case of strong dependencies. DepSMUCE and JUSD show a good and similar performance if the error process is an MA(4)-process and the corresponding parameters are reasonably large.

Finally, we consider an example where the error process in model (1.1) is a stationary and causal ARMA$(2,6)$-process defined by

$$\varepsilon_i = 0.75\varepsilon_{i-1} - 0.5\varepsilon_{i-2} + \eta_i + 0.8\eta_{i-1} + 0.7\eta_{i-2} + 0.6\eta_{i-3} + 0.5\eta_{i-4} + 0.4\eta_{i-5} + 0.3\eta_{i-6}, \tag{4.6}$$

where $\{\eta_i\}_{i\in\mathbb{Z}}$ is a sequence of independent standard normal distributed random variables. We consider again a model with $K^* = 5$ change points located as described in (4.2) with the corresponding function intensities

$$(\theta_0^*, \theta_1^*, \theta_2^*, \theta_3^*, \theta_4^*, \theta_5^*) = (0, 5, 1, 8, 1, -2) \tag{4.7}$$

(see Figure 1). The data driven procedure from the R-package *dbacf* [Tecuapetla-Gómez (2015)] now leads to ambiguous results because there is no correct m to estimate. Table 6 shows the estimated numbers of change points. While at level $\alpha = 0.5$ DepSMUCE correctly estimates $K^* = 5$ in more than 93% of the cases, JUSD mostly overestimates K^*. From the right part of Table 5 we also observe that K^* is estimated more precisely by DepSMUCE than by JUSD with smaller MSE and MAE. As in the MA(4)-example, SMUCE in general includes a large amount of false positives. Finally, these results are reflected in Figure 3, where we show the histograms of estimated change points. For the ARMA$(2,6)$ error process DepSMUCE yields substantially better results than JUSD.

$\hat{K} - K^*$	≤ -3	-2	-1	0	$+1$	$+2$	$\geq +3$
SMUCE(0.1)	0.000	0.000	0.000	0.000	0.000	0.000	1.000
SMUCE(0.5)	0.000	0.000	0.000	0.000	0.000	0.000	1.000
SMUCE(0.9)	0.000	0.000	0.000	0.000	0.000	0.000	1.000
DepSMUCE(0.1)	0.001	0.061	0.391	0.547	0.000	0.000	0.000
DepSMUCE(0.5)	0.000	0.001	0.049	0.937	0.013	0.000	0.000
DepSMUCE(0.9)	0.000	0.000	0.005	0.892	0.096	0.007	0.000
JUSD(0.1)	0.055	0.144	0.203	0.292	0.099	0.077	0.129
JUSD(0.5)	0.005	0.022	0.087	0.447	0.077	0.066	0.296
JUSD(0.9)	0.001	0.002	0.012	0.444	0.067	0.033	0.441

Table 6: *Proportion of estimated numbers of change points (the true number of change points is $K^* = 5$) in model* (1.1) *with step function defined by* (4.2) *and* (4.7) *and an ARMA* $(2,6)$ *error process defined in* (4.6).

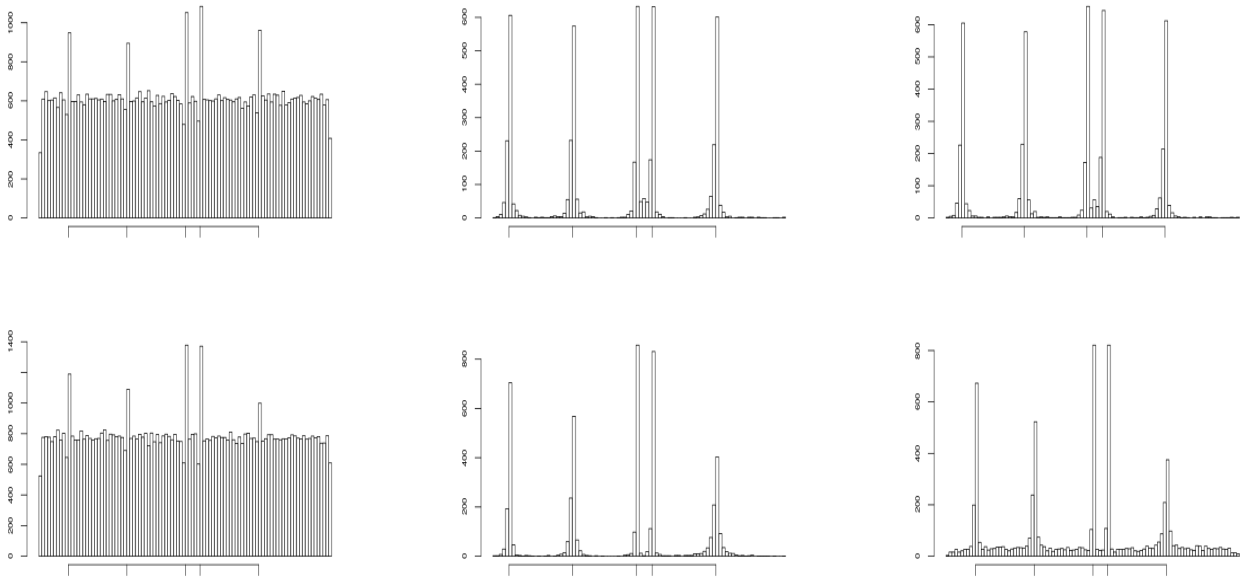

Figure 3: *Histograms of estimated change point locations for different estimators. Upper row: MA*(4) *error process. Lower row: ARMA*$(2, 6)$ *error process. Left columun: SMUCE. Middle columun: DepSMUCE. Right columun: JUSD. The "true" change points are located at* 101*,* 301*,* 501*,* 551*, and* 751*.*

5 Proofs

5.1 Proof of Theorem 3.2

The proof essentially proceeds in two steps. First we will prove an analog of the result for the statistic

$$V_{n,\sigma_\star}(Y,\vartheta) = \max_{0\le k\le K} \max_{\substack{n\tau_k\le i\le j<n\tau_{k+1}\\ j-i+1\ge nc_n}} \left\{ \frac{1}{\sigma_\star}\sqrt{j-i+1}\left|\overline{Y}_i^j - \theta_k\right| - \sqrt{2\log\frac{en}{j-i+1}}\right\},$$

which is defined as V_n using the known long run variance. This result is essentially based on a Gaussian approximation via assumption (A3). In a second step we use Proposition 3.1, for which assumptions (A1) and (A2) are needed, to show that the error caused by the estimation of the long run variance is negligible. The proof of the proposition is given at the end of this section.

Theorem 5.1 *Consider the nonparametric regression model* (1.1) *and assume that* $\|\varepsilon_i\|_3 < \infty$*.*

If assumption (A3) holds for some $\gamma > 1/2$ and $c_n \to 0$ fulfils (3.1)*, then we have*

$$V_{n,\sigma_\star}(Y,\vartheta^*) \xrightarrow{\mathcal{D}} \max_{0\le k\le K^*} \sup_{\tau_k^*\le s<t\le\tau_{k+1}^*} \left\{ \frac{|B(t)-B(s)|}{\sqrt{t-s}} - \sqrt{2\log\frac{e}{t-s}} \right\} \text{ as } n\to\infty,$$

where $\{B(t)\}_{t\in[0,1]}$ denotes a standard Brownian motion.

Proof of Theorem 5.1 : By (A3), the assumptions of Theorem 1 in Wu and Zhou (2011) are fulfilled. It therefore follows that on a richer probability space $(\check{\Omega},\check{\mathcal{A}},\check{\mathbb{P}})$, there exists a process $\{\check{S}_i\}_{i=1}^n$ and a centered Gaussian process $\{\check{G}_i\}_{i=1}^n$ with independent increments such that

$$\Big(\sum_{\ell=1}^{i}\varepsilon_\ell\Big)_{i=1}^n \overset{\mathcal{D}}{=} \big(\check{S}_i\big)_{i=1}^n \text{ and } \max_{1\le i\le n}\big|\check{S}_i-\check{G}_i\big| = \mathcal{O}_{\check{\mathbb{P}}}(\tau_n), \tag{5.1}$$

where

$$\tau_n = n^{(1+2\gamma)/(1+6\gamma)}(\log n)^{8\gamma/(1+6\gamma)}.$$

Moreover, again on a richer probability space $(\hat{\Omega},\hat{\mathcal{A}},\hat{\mathbb{P}})$, there exists another Gaussian process $\{\hat{G}_i\}_{i=1}^n$ and i.i.d. random variables $U_\ell \sim \mathcal{N}(0,\sigma_\star^2)$ such that

$$\big(\check{G}_i\big)_{i=1}^n \overset{\mathcal{D}}{=} \big(\hat{G}_i\big)_{i=1}^n \text{ and } \max_{1\le i\le n}\Big|\hat{G}_i-\sum_{\ell=1}^{i}U_\ell\Big| = \mathcal{O}_{\hat{\mathbb{P}}}(\tau_n). \tag{5.2}$$

By (5.1), it follows that $V_{n,\sigma_\star}(Y,\vartheta^*) \overset{\mathcal{D}}{=} \check{V}_{n,\sigma_\star}(Y,\vartheta^*)$, where

$$\check{V}_{n,\sigma_\star}(Y,\vartheta^*) = \max_{0\le k\le K^*}\max_{\substack{n\tau_k^*\le i\le j<n\tau_{k+1}^*\\ j-i+1\ge nc_n}} \left\{ \frac{1}{\sigma_\star\sqrt{j-i+1}}\left|\check{S}_j-\check{S}_{i-1}\right| - \sqrt{2\log\frac{en}{j-i+1}} \right\}.$$

By (3.1) we have $\tau_n = \mathrm{o}(\sqrt{nc_n})$, and a further application of (5.1) and the triangle inequality yield

$$\check{V}_{n,\sigma_\star}(Y,\vartheta^*) = \max_{0\le k\le K^*}\max_{\substack{n\tau_k^*\le i\le j<n\tau_{k+1}^*\\ j-i+1\ge nc_n}} \left\{ \frac{1}{\sigma_\star\sqrt{j-i+1}}\left|\check{G}_j-\check{G}_{i-1}\right| - \sqrt{2\log\frac{en}{j-i+1}} \right\} + \mathrm{o}_{\check{\mathbb{P}}}(1).$$

By (5.2), the first random variable on the right hand side has the same distribution as

$$\hat{V}_{n,\sigma_\star}(Y,\vartheta^*) = \max_{0\le k\le K^*}\max_{\substack{n\tau_k^*\le i\le j<n\tau_{k+1}^*\\ j-i+1\ge nc_n}} \left\{ \frac{1}{\sigma_\star\sqrt{j-i+1}}\left|\hat{G}_j-\hat{G}_{i-1}\right| - \sqrt{2\log\frac{en}{j-i+1}} \right\}.$$

With the same arguments as given above, we obtain

$$\hat{V}_{n,\sigma_\star}(Y,\vartheta^*) = \max_{0\le k\le K^*}\max_{\substack{n\tau_k^*\le i\le j<n\tau_{k+1}^*\\ j-i+1\ge nc_n}} \left\{ \frac{1}{\sigma_\star}\sqrt{j-i+1}\left|\overline{U}_i^j\right| - \sqrt{2\log\frac{en}{j-i+1}} \right\} + \mathrm{o}_{\hat{\mathbb{P}}}(1).$$

Note that

$$\max_{0\le k\le K^*}\max_{\substack{n\tau_k^*\le i\le j<n\tau_{k+1}^*\\ j-i+1\ge nc_n}}\left\{\frac{1}{\sigma_\star}\sqrt{j-i+1}\left|\overline{U}_i^j\right|-\sqrt{2\log\frac{en}{j-i+1}}\right\}$$
$$\stackrel{\mathcal{D}}{=}\max_{0\le k\le K^*}\max_{\substack{n\tau_k^*\le i\le j<n\tau_{k+1}^*\\ j-i+1\ge nc_n}}\left\{\sqrt{j-i+1}\left|\overline{Z}_i^j\right|-\sqrt{2\log\frac{en}{j-i+1}}\right\},$$

where $Z_1,\dots,Z_n\sim\mathcal{N}(0,1)$ are i.i.d. The assertion now follows with the same arguments as given in the proof of Theorem 1 in Frick et al. (2014). □

Proof of Theorem 3.2 : For the sake of clarity, we will denote the statistic V_n in (2.2) by $V_{n,\hat{\sigma}_\star}$ to emphasize its dependence on the estimator $\hat{\sigma}_\star^2$ of the long run variance. Considering the proof of Theorem 5.1, it suffices to show that

$$V_{n,\sigma_\star}(Y,\vartheta^*)-V_{n,\hat{\sigma}_\star}(Y,\vartheta^*)=\mathrm{o}_{\mathbb{P}}(1). \tag{5.3}$$

A straightforward application of the triangle inequality yields

$$|V_{n,\sigma_\star}(Y,\vartheta^*)-V_{n,\hat{\sigma}_\star}(Y,\vartheta^*)|\le\left|\frac{1}{\sigma_\star}-\frac{1}{\hat{\sigma}_\star}\right|\max_{0\le k\le K^*}\max_{\substack{n\tau_k^*\le i\le j<n\tau_{k+1}^*\\ j-i+1\ge nc_n}}\left|\sqrt{j-i+1}\left(\overline{Y}_i^j-\theta_k^*\right)\right|.$$

We use again the Gaussian approximation result from Theorem 5.1 and obtain

$$D_n:=\max_{\substack{1\le i\le j\le n\\ j-i+1\ge nc_n}}\left|\sqrt{j-i+1}\left(\overline{Y}_i^j-\mathbb{E}[\overline{Y}_i^j]\right)\right|\stackrel{\mathcal{D}}{=}\max_{\substack{1\le i\le j\le n\\ j-i+1\ge nc_n}}\left\{\frac{1}{\sqrt{j-i+1}}\left|\check{S}_j-\check{S}_{i-1}\right|\right\}$$

as well as

$$\max_{\substack{1\le i\le j\le n\\ j-i+1\ge nc_n}}\left\{\frac{1}{\sqrt{j-i+1}}\left|\check{S}_j-\check{S}_{i-1}\right|\right\}=\max_{\substack{1\le i\le j\le n\\ j-i+1\ge nc_n}}\left\{\frac{1}{\sqrt{j-i+1}}\left|\check{G}_j-\check{G}_{i-1}\right|\right\}+\mathrm{o}_{\check{\mathbb{P}}}(1).$$

Furthermore, it holds that

$$\max_{\substack{1\le i\le j\le n\\ j-i+1\ge nc_n}}\left\{\frac{1}{\sqrt{j-i+1}}\left|\check{G}_j-\check{G}_{i-1}\right|\right\}\stackrel{\mathcal{D}}{=}\max_{\substack{1\le i\le j\le n\\ j-i+1\ge nc_n}}\left\{\frac{1}{\sqrt{j-i+1}}\left|\hat{G}_j-\hat{G}_{i-1}\right|\right\}$$

and

$$\max_{\substack{1\le i\le j\le n\\ j-i+1\ge nc_n}}\left\{\frac{1}{\sqrt{j-i+1}}\left|\hat{G}_j-\hat{G}_{i-1}\right|\right\}=\max_{\substack{1\le i\le j\le n\\ j-i+1\ge nc_n}}\left\{\sqrt{j-i+1}\left|\overline{U}_i^j\right|\right\}+\mathrm{o}_{\hat{\mathbb{P}}}(1).$$

From Theorem 1 in Shao (1995) it follows that

$$\max_{\substack{1\le i\le j\le n\\ j-i+1\ge nc_n}}\left\{\sqrt{j-i+1}\left|\overline{U}_i^j\right|\right\}\le\max_{1\le i\le j\le n}\left\{\sqrt{j-i+1}\left|\overline{U}_i^j\right|\right\}=\mathcal{O}(\sqrt{\log n})\quad\text{a.s.}$$

In combination with Proposition 3.1, this yields (5.3). □

5.2 Proof of Theorem 3.4

Note that by definition of $\hat{K}$, we have

$$\hat{K}(V_n, q_n) < K^* \quad \Longleftrightarrow \quad \exists \vartheta \in \mathcal{S}_n \text{ with } |J(\vartheta)| < K^* \text{ such that } V_n(Y, \vartheta) \leq q_n$$

[see Frick et al. (2014)]. It therefore suffices to show that the probability of the existence of a candidate function $\vartheta \in \mathcal{S}_n$ having less than K^* change points and fulfilling $V_n(Y, \vartheta) \leq q_n$ converges to 0.

We will again first prove an analog of the result for the statistic $V_{n,\sigma_\star}$, where the estimator $\hat{\sigma}_\star^2$ is replaced by the long run variance.

Theorem 5.2 *Under the same assumptions as in Theorem 5.1, assume that* $\{q_n\}_{n\in\mathbb{N}}$ *is a sequence fulfilling* (3.4)*. Then it follows that*

$$\lim_{n\to\infty} \mathbb{P}\left(\hat{K}(V_{n,\sigma_\star}, q_n) < K^*\right) = 0.$$

Proof of Theorem 5.2 : We proceed similarly as in the proof of Theorem 7.10 in Frick et al. (2014). Set

$$\lambda := \inf_{0\leq k\leq K^*} \left|\tau_{k+1}^* - \tau_k^*\right|, \quad \beta := \inf_{1\leq k\leq K^*} \left|\theta_k^* - \theta_{k-1}^*\right|,$$

and define the K^* disjoint intervals

$$I_i := \left[\tau_i^* - \frac{\lambda}{2}, \tau_i^* + \frac{\lambda}{2}\right), \quad i = 1, \ldots, K^*.$$

Moreover, define $\theta_i^+ := \max\{\theta_{i-1}^*, \theta_i^*\}$, $\theta_i^- := \min\{\theta_{i-1}^*, \theta_i^*\}$, $I_i^+ := \{t \in I_i : \vartheta^*(t) = \theta_i^+\}$, and $I_i^- := \{t \in I_i : \vartheta^*(t) = \theta_i^-\}$. Note that $|I_i^+| = |I_i^-| = \lambda/2$. In particular, since $\{c_n\}_{n\in\mathbb{N}}$ is a null sequence, it holds that $|I_i^+| \geq c_n$ and $|I_i^-| \geq c_n$ for any $n \in \mathbb{N}$ large enough.

Any candidate function with $K < K^*$ change points must be constant on at least one of the disjoint intervals I_i. Therefore we get

$$\begin{aligned}\mathbb{P}\left(\hat{K}(V_{n,\sigma_\star}, q_n) < K^*\right) \leq &\sum_{i=1}^{K^*} \mathbb{P}\left(\exists \theta \leq \theta_i^+ - \frac{\beta}{2} \,:\, \frac{1}{\sigma_\star}\sqrt{\frac{n\lambda}{2}}\left|\overline{Y}_{I_i^+} - \theta\right| - \sqrt{2\log\frac{2e}{\lambda}} \leq q_n\right) \\ &+ \sum_{i=1}^{K^*} \mathbb{P}\left(\exists \theta \geq \theta_i^- + \frac{\beta}{2} \,:\, \frac{1}{\sigma_\star}\sqrt{\frac{n\lambda}{2}}\left|\overline{Y}_{I_i^-} - \theta\right| - \sqrt{2\log\frac{2e}{\lambda}} \leq q_n\right).\end{aligned}$$

All of these summands can be dealt with analogously, which is why we will restrict ourselves to the second probability and the case $i = 1$. Without loss of generality, assume that $I_1^- = [\tau_1^* - \lambda/2, \tau_1^*)$. It follows easily that the term of interest is bounded from above by

$$\mathbb{P}\left(\frac{1}{\sigma_\star}\sqrt{\frac{n\lambda}{2}}\left|\overline{\varepsilon}_{n\tau_1^* - \frac{n\lambda}{2}}^{n\tau_1^*-1} - \frac{\beta}{2}\right| - \sqrt{2\log\frac{2e}{\lambda}} \leq q_n\right) + \mathbb{P}\left(\overline{\varepsilon}_{n\tau_1^* - \frac{n\lambda}{2}}^{n\tau_1^*-1} > \frac{\beta}{2}\right). \tag{5.4}$$

Since $\{\varepsilon_i\}_{i\in\mathbb{Z}}$ is mean-ergodic, the second probability in (5.4) converges to 0. Concerning the first probability in (5.4), note that with exactly the same Gaussian approximation arguments as given in the proof of Theorem 5.1, it suffices to show that

$$\hat{\mathbb{P}}\left(\frac{1}{\sigma_\star}\sqrt{\frac{n\lambda}{2}}\left|\overline{U}_{n\tau_1^*-\frac{n\lambda}{2}}^{n\tau_1^*-1}-\frac{\beta}{2}\right|-\sqrt{2\log\frac{2e}{\lambda}}\leq q_n\right)=\mathrm{o}(1), \tag{5.5}$$

where $U_1,\ldots,U_n\sim\mathcal{N}(0,\sigma_\star^2)$ are i.i.d. and defined on a richer probability space $(\hat{\Omega},\hat{\mathcal{A}},\hat{\mathbb{P}})$. From Theorem 7.10 and Lemma 7.11 in Frick et al. (2014), it follows that the probability in (5.5) is upper bounded by $e^{-\frac{1}{64\sigma_\star^2}n\lambda\beta^2+\frac{1}{2}(q_n+\sqrt{2\log\frac{2e}{\lambda}})^2}$. By assumption (3.4), this expression vanishes as n tends to ∞. □

Proof of Theorem 3.4 : With exactly the same arguments as given in the proof of Theorem 5.2, it suffices to show that

$$\lim_{n\to\infty}\mathbb{P}\left(\frac{1}{\hat{\sigma}_\star}\sqrt{\frac{n\lambda}{2}}\left|\overline{\varepsilon}_{n\tau_1^*-\frac{n\lambda}{2}}^{n\tau_1^*-1}-\frac{\beta}{2}\right|-\sqrt{2\log\frac{2e}{\lambda}}\leq q_n\right)=0.$$

Set

$$X_n:=\frac{1}{\sigma_\star}\sqrt{\frac{n\lambda}{2}}\left|\overline{\varepsilon}_{n\tau_1^*-\frac{n\lambda}{2}}^{n\tau_1^*-1}-\frac{\beta}{2}\right| \quad\text{and}\quad Y_n:=\frac{1}{\hat{\sigma}_\star}\sqrt{\frac{n\lambda}{2}}\left|\overline{\varepsilon}_{n\tau_1^*-\frac{n\lambda}{2}}^{n\tau_1^*-1}-\frac{\beta}{2}\right|.$$

Let $\delta>0$ be arbitrary and note that

$$\begin{aligned}\mathbb{P}\left(Y_n\leq q_n+\sqrt{2\log\frac{2e}{\lambda}}\right)&=\mathbb{P}\left(Y_n\leq q_n+\sqrt{2\log\frac{2e}{\lambda}}\,,\,\left|\frac{Y_n}{X_n}-1\right|>\delta\right)\\&\quad+\mathbb{P}\left(Y_n\leq q_n+\sqrt{2\log\frac{2e}{\lambda}}\,,\,\left|\frac{Y_n}{X_n}-1\right|\leq\delta\right).\end{aligned}$$

By Proposition 3.1, the first probability converges to 0. Concerning the second probability, it holds that

$$\mathbb{P}\left(Y_n\leq q_n+\sqrt{2\log\frac{2e}{\lambda}}\,,\,\left|\frac{Y_n}{X_n}-1\right|\leq\delta\right)\leq\mathbb{P}\left((1-\delta)\,X_n\leq q_n+\sqrt{2\log\frac{2e}{\lambda}}\right).$$

With the arguments given in the proof of Theorem 5.2, the expression on the right hand side is bounded by $e^{-\frac{1}{64\sigma_\star^2}n\lambda\beta^2+\frac{1}{2}\left((q_n+\sqrt{2\log\frac{2e}{\lambda}})/(1-\delta)\right)^2}+\mathrm{o}(1)$, which converges to 0 by (3.4) for a fixed $\delta>0$. □

5.3 Proof of Theorem 3.6

We will prove a corresponding statement for the statistic $V_{n,\sigma_\star}$ where again the estimator $\hat{\sigma}_\star^2$ is replaced by the long run variance, that is

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{\vartheta\in\mathcal{C}(V_{n,\sigma_\star},q_n)} \max_{\tau^*\in J(\vartheta^*)} \min_{\tau\in J(\vartheta)} |\tau^*-\tau| > c_n \right) = 0.$$

For a proof of this statement we define the value β and the intervals J_i, J_i^-, and J_i^+ as in the proof of Theorem 5.2 by replacing $\lambda/2$ with c_n and then using the letter J instead of I. Any candidate function $\vartheta \in \mathcal{S}_n$ with

$$\max_{\tau^*\in J(\vartheta^*)} \min_{\tau\in J(\vartheta)} |\tau^*-\tau| > c_n$$

must be constant on at least one of the disjoint intervals J_i. Assume without loss of generality that $J_1^- = [\tau_1^* - c_n, \tau_1^*)$. With the same arguments as given in the proof of Theorem 5.2, it suffices to show that

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{\sigma_\star}\sqrt{nc_n} \left| \overline{\varepsilon}_{n\tau_1^*-nc_n}^{n\tau_1^*-1} - \frac{\beta}{2} \right| - \sqrt{2\log\frac{e}{c_n}} \le q_n \right) = 0. \tag{5.6}$$

Theorem 7.10 and Lemma 7.11 in Frick et al. (2014) in combination with the Gaussian approximation arguments from the proof of Theorem 5.1 yield that the probability in (5.6) is bounded by $e^{-\frac{1}{32\sigma_\star^2}nc_n\beta^2+\frac{1}{2}(q_n+\sqrt{2\log\frac{e}{c_n}})^2} + \mathrm{o}(1)$, which converges to 0 by (3.5).

For the proof of Theorem 3.6 assume again that $J_1^- = [\tau_1^* - c_n, \tau_1^*)$ and note that the same arguments show that the assertion follows from the statement

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{\hat{\sigma}_\star}\sqrt{nc_n} \left| \overline{\varepsilon}_{n\tau_1^*-nc_n}^{n\tau_1^*-1} - \frac{\beta}{2} \right| - \sqrt{2\log\frac{e}{c_n}} \le q_n \right) = 0.$$

The proof thus works exactly as the proof of Theorem 3.4. □

5.4 Proof of Proposition 3.1

Note that it suffices to show that

$$\mathbb{E}\left[\left(\hat{\sigma}_\star^2 - \sigma_\star^2 \right)^2 \right] = \mathcal{O}(n^{-2/3}).$$

We proceed as in the proof of Theorem 3 in Wu and Zhao (2007). Moreover, by assumption (A2), we can apply Lemma 4 and Lemma 5 from Wu and Zhao (2007). For $2 \le i \le m_n$ we define

$$W_{ik_n} := \sum_{j=(i-1)k_n+1}^{ik_n} \varepsilon_j - \sum_{j=(i-2)k_n+1}^{(i-1)k_n} \varepsilon_j$$

and

$$r_{ik_n} := \sum_{j=(i-1)k_n+1}^{ik_n} \vartheta^*\left(\frac{j}{n}\right) - \sum_{j=(i-2)k_n+1}^{(i-1)k_n} \vartheta^*\left(\frac{j}{n}\right).$$

By Lemma 4 in Wu and Zhao (2007) it then follows that

$$\begin{aligned}\mathbb{E}\left[\left(\hat{\sigma}_\star^2 - \sigma_\star^2\right)^2\right] &= \Big|\Big|\frac{1}{2k_n(m_n-1)}\sum_{i=2}^{m_n}(W_{ik_n}+r_{ik_n})^2 - \sigma_\star^2\Big|\Big|_2^2 \\ &\leq \Big|\Big|\frac{1}{2k_n(m_n-1)}\sum_{i=2}^{m_n}\left[(W_{ik_n}+r_{ik_n})^2 - ||W_{ik_n}||_2^2\right]\Big|\Big|_2^2 + \mathcal{O}(k_n^{-2}).\end{aligned}$$

Note that

$$\begin{aligned}\Big|\Big|\frac{1}{2k_n(m_n-1)}\sum_{i=2}^{m_n}\left[(W_{ik_n}+r_{ik_n})^2 - ||W_{ik_n}||_2^2\right]\Big|\Big|_2^2 &\leq \frac{1}{4k_n^2(m_n-1)^2}\Big|\Big|\sum_{i=2}^{m_n}\left[(W_{ik_n}+r_{ik_n})^2 - W_{ik_n}^2\right]\Big|\Big|_2^2 \\ &\quad + \frac{1}{4k_n^2(m_n-1)^2}\Big|\Big|\sum_{i=2}^{m_n}W_{ik_n}^2 - (m_n-1)||W_{2k_n}||_2^2\Big|\Big|_2^2.\end{aligned}$$

By Lemma 5 in Wu and Zhao (2007), the second term is of the order $\mathcal{O}(m_n^{-1})$. We therefore need to deal with

$$\frac{1}{4k_n^2(m_n-1)^2}\Big|\Big|\sum_{i=2}^{m_n}\left[2W_{ik_n}r_{ik_n} + r_{ik_n}^2\right]\Big|\Big|_2^2.$$

Lemma 4 in Wu and Zhao (2007) gives that $||W_{ik_n}||_2 = \mathcal{O}(\sqrt{k_n})$ uniformly over $i = 2, \ldots, m_n$. Moreover, it holds that $r_{ik_n} = \mathcal{O}(k_n)$ uniformly over $i = 2, \ldots, m_n$. Note that since ϑ^* is piecewise constant with $K^* < \infty$ change points, the set

$$\left\{i \in \{2, \ldots, m_n\} \,\middle|\, r_{ik_n} \neq 0\right\}$$

contains a finite number of elements, independently of $n \in \mathbb{N}$. Therefore, it follows that

$$\Big|\Big|\sum_{i=2}^{m_n}\left[2W_{ik_n}r_{ik_n} + r_{ik_n}^2\right]\Big|\Big|_2 = \mathcal{O}(k_n^2),$$

which yields that

$$\frac{1}{4k_n^2(m_n-1)^2}\Big|\Big|\sum_{i=2}^{m_n}\left[2W_{ik_n}r_{ik_n} + r_{ik_n}^2\right]\Big|\Big|_2^2 = \mathcal{O}\left(k_n^2 m_n^{-2}\right).$$

Since $k_n \asymp n^{1/3}$, the claim follows. □

Acknowledgements This work has been supported in part by the Collaborative Research Center "Statistical modeling of nonlinear dynamic processes" (SFB 823, Teilprojekt A1, C1) of the German Research Foundation (DFG). The authors would like to thank Wei Biao Wu for helpful discussions regarding physical dependence and Thomas Hotz, Axel Munk and Florian Pein for helpful discussions about the algorithmic aspects of SMUCE.

References

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

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *J. Appl. Econmetr.*, 18:1–22.

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

Chakar, S., Lebarbier, E., Lévy-Leduc, C., and Robin, S. (2017). A robust approach for estimating change-points in the mean of an AR(1) process. *Bernoulli*, 23:1408–1447.

Cho, H. and Fryzlewicz, P. (2015). Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *J. R. Statist. Soc. B*, 77:475–507.

Ciuperca, G. (2011). A general criterion to determine the number of change-points. *Statist. Probab. Lett.*, 81:1267–1275.

Ciuperca, G. (2014). Model selection by LASSO methods in a change-point model. *Stat. Papers*, 55:349–374.

Davis, R. A., Lee, T. C. M., and Rodriguez-Yam, G. A. (2006). Structural break estimation for nonstationary time series models. *J. Am. Statist. Ass.*, 101:223–239.

Dette, H., Munk, A., and Wagner, T. (1998). Estimating the variance in nonparametric regression - what is a reasonable choice? *J. Roy. Stat. Soc. B*, 60:751–764.

Frick, K., Munk, A., and Sieling, H. (2014). Multiscale change point inference. *Journal of the Royal Statistical Society, Ser. B*, 76(3):495–580.

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

Hall, P., Kay, J. W., and Titterington, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, 77:521–528.

Harchaoui, Z. and Lévy-Leduc, C. (2010). Multiple change-point estimation with a total variation penalty. *J. Am. Statist. Ass.*, 105:1480–1493.

Haynes, K., Fearnhead, P., and Eckley, I. A. (2017). A computationally efficient nonparametric approach for changepoint detection. *Stat. Comput.*, 27:1293–1305.

Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *J. Am. Statist. Ass.*, 107:1590–1598.

Kolaczyk, E. D. and Nowak, R. D. (2005). Multiscale generalised linear models for nonparametric function estimation. *Biometrika*, 92:119–133.

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

Lavielle, M. and Moulines, E. (2000). Least-squares estimation of an unknown number of shifts in a time series. *J. Time Series Anal.*, 21:33–59.

Li, H., Guo, Q., and Munk, A. (2018). Multiscale change-point segmentation: Beyond step functions. *arXiv:1708.03942*.

Li, H., Munk, A., and Sieling, H. (2016). FDR-control in multiscale change-point segmentation. *Electron. J. Statist.*, 10:918–959.

Matteson, D. S. and James, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *J. Am. Statist. Ass.*, 109:334–345.

Pein, F., Hotz, T., Sieling, H., and Aspelmeier, T. (2017a). *stepR: Multiscale change-point inference*. R package version 2.0-1.

Pein, F., Sieling, H., and Munk, A. (2017b). Heterogeneous change point inference. *Journal of the Royal Statistical Society, Ser. B*, 79(4):1207–1227.

Preuss, P., Puchstein, R., and Dette, H. (2015). Detection of multiple structural breaks in multivariate time series. *J. Am. Statist. Ass.*, 110:654–668.

Shao, Q.-M. (1995). On a conjecture of révész. *Proceedings of the American Mathematical Society*, 123(2):575–582.

Tecuapetla-Gómez, I. (2015). *dbacf: Autocovariance Estimation via Difference-Based Methods*. R package version 0.0.0.9000.

Tecuapetla-Gómez, I. and Munk, A. (2017). Autocovariance estimation in regression with a discontinuous signal and m-dependent errors: A difference-based approach. *Scandinavian Journal of Statistics*, 44(2):346–368.

Wu, B. W. (2005). Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences USA*, 102:14150–14154.

Wu, B. W. (2011). Asymptotic theory for stationary processes. *Statistics and Its Interface*, 4:207–226.

Wu, B. W. and Zhao, Z. (2007). Inference of trends in time series. *Journal of the Royal Statistical Society, Ser. B*, 69(3):391–410.

Wu, B. W. and Zhou, Z. (2011). Gaussian approximations for non-stationary multiple time series. *Statistica Sinica*, 21:1397–1413.

Wu, W. B. and Phoumaradi, M. (2009). Banding sample autocovariance matrices of stationary processes. *Statistica Sinica*, 19:1755–1768.

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

Yau, C. Y. and Zhao, Z. (2016). Inference for multiple change points in time series via likelihood ratio scan statistics. *J. Roy. Stat. Soc. B*, 78:895–916.