Modified sequential change point procedures based on estimating functions[*],[†]

Claudia Kirch[‡]

*Institute of Mathematical Stochastics, Department of Mathematics, Otto-von-Guericke
University, Magdeburg, Germany
Center for Behavioral Brain Sciences (CBBS), Magdeburg, Germany
e-mail:* claudia.kirch@ovgu.de
url: http://www.imst1.ovgu.de/imst1/en/

Silke Weber

*Karlsruhe Institute of Technology (KIT), Institute of Stochastics, Englerstr. 2, 76131
Karlsruhe, Germany
e-mail:* silke.weber@kit.edu

Abstract: A large class of sequential change point tests are based on estimating functions where estimation is computationally efficient as (possibly numeric) optimization is restricted to an initial estimation. This includes examples as diverse as mean changes, linear or non-linear autoregressive and binary models. While the standard cumulative-sum-detector (CUSUM) has recently been considered in this general setup, we consider several modifications that have faster detection rates in particular if changes occur late in the monitoring period. More presicely, we use three different types of detector statistics based on partial sums of a monitoring function, namely the modified moving-sum-statistic (mMOSUM), Page's cumulative-sum-statistic (Page-CUSUM) and the standard moving-sum-statistic (MOSUM). The statistics only differ in the number of observations included in the partial sum. The mMOSUM uses a bandwidth parameter which multiplicatively scales the lower bound of the moving sum. The MOSUM uses a constant bandwidth parameter, while Page-CUSUM chooses the maximum over all possible lower bounds for the partial sums. So far, the first two schemes have only been studied in a linear model, the MOSUM only for a mean change.

We develop the asymptotics under the null hypothesis and alternatives under mild regularity conditions for each test statistic, which include the existing theory but also many new examples. In a simulation study we compare all four types of test procedures in terms of their size, power and run length. Additionally we illustrate their behavior by applications to exchange rate data as well as the Boston homicide data.

MSC 2010 subject classifications: 62L10, 62G10, 62M10.
Keywords and phrases: Change point analysis, sequential test, online monitoring, regression, integer-valued time series.

1. Introduction

Despite a relatively long tradition in statistics, change point analysis is a very active field and has become increasingly popular in the last years due to its importance in many areas where data is collected over time. In addition to a-posteriori change point methods, there has been a recent interest in sequential methodology due to the fact that an increasing number of data sets are collected automatically or without significant costs such that the observations arrive steadily. Examples include financial data sets, e.g., in risk management (Andreou and Ghysels (2006)) or CAPM models (Aue et al. (2011)) as well as medical data sets, e.g., monitoring intensive care patients (Fried and Imhoff (2004)).

Chu et al. (1996) introduced a new way of sequential testing, which allows to control the asymptotic α-error if no changes occur while having asymptotic power one under alternatives, which has then be pursued by others e.g. Horváth et al. (2004), Aue et al. (2006), Horváth et al. (2008), Aue et al. (2008), Aue and Horváth (2004), Fremdt (2014) and Hušková and Koubková (2005) (see Section 1.2 below for more details). The key assumption is the existence of a historic data set which is used for initial estimation before monitoring starts. This is usually given in applications as some data collection must have taken place before one can safely build or monitor a model. This approach allows for nonparametric inference by means of asymptotics by letting the length of the historic data set grow to infinity even if a possibly infinite observation horizon is used. Nonparametric inference is meant in the sense that e.g. the stochastic structure of the innovation process in a regression model is not completely specified.

As recently pointed out by Kirch and Kamgaing (2015) most of these statistics can be written by means of estimating and monitoring functions, which then provide a unifying framework for the derivation of the asymptotic results. Kirch and Kamgaing (2015) derive regularity conditions under which they prove limit results for the standard CUSUM monitoring scheme as originally proposed by Chu et al. (1996), that has been used in most follow-up works for different change point scenarios, e.g. Horváth et al. (2004), Aue et al. (2006), Aue and Horváth (2004) and Hušková and Koubková (2005).

The main disadvantage of the standard sequential CUSUM statistic is the fact that the detection time can be rather long if changes occur late in the monitoring period. This is due to the fact that effectively a (properly scaled) two-sample test is applied comparing the historic data set with the data set after monitoring starts up to the most recent observation. Thus for a late change many 'null observations' contaminate the second data set so that more 'alternative' observations need to be collected before significance is reached. This is why, recently, alternative monitoring schemes based only on more recent observations have been proposed in the literature: The mMOSUM in a linear model setup in Chen and Tian (2010) (where their null limit distribution is not correct), the Page-CUSUM in a linear model setup in Fremdt (2014) and the standard MOSUM for the location model in Horváth et al. (2008) and Aue et al. (2008).

In this paper, we generalize the latter monitoring schemes to the general framework of estimating functions thus greatly extending their range of applications to examples as diverse as the non-linear autoregressive or the binary time series model. Furthermore, we investigate the differences in terms of size, power and run length in a simulation study and illustrate the behavior on two data sets.

The paper is organized as follows: First, the general setup and the new statistics are introduced, namely the mMOSUM, the Page-CUSUM and the MOSUM statistic. Examples for change point problems that fall into that framework are given in Section 1.2. In Section 2 we develop the limit distributions for these statistics in the general setting under the null hypothesis and mild assumptions. We first focus on the mMOSUM and the Page-CUSUM statistic before developing the asymptotics for the MOSUM statistic under slightly different assumptions. In Subsection 2.3 we show that the different procedures have asymptotic power one under very general conditions. Finally, in Section 3 we compare the different types of statistics, including the CUSUM statistic in an extensive simulation study by means of their empirical size, power and run lengths in different scenarios before giving two data examples.

1.1. Monitoring schemes based on estimating functions

As already mentioned we assume the existence of a historic training data set $X_1, \ldots, X_m$ with no change points. Based on this initial observations we estimate the unknown parameter $\theta_0 \subset \mathbb{R}^d$ by means of an estimating function, i.e. the estimator $\widehat{\theta}_m$ is obtained as the zero of the following sum:

$$\sum_{t=1}^{m} G(X_t, \widehat{\theta}_m) = 0, \tag{1}$$

where $X_t, t = 1, ..., m$, are the historical data and the estimating function G takes values in $\mathbb{R}^d$. All vectors given in this work are column vectors.

The simplest example is the estimating function $G(x, \theta) = (x - \theta)$ in a location model which leads to the mean (as estimator for the expectation), i.e. $\widehat{\theta}_m = \bar{X}_m$. In a linear model with $Y_t = \theta^T Z_t + e_t$, we consider $X_t^T = (Y_t, Z_t^T)$ as well as $G((y, z), \theta) = z(y - \theta^T z)$, which yields the usual least squares estimator in this context. Similarly, in an autoregressive linear model Z_t is given by the lagged observations $(Y_{t-1}, \ldots, Y_{t-p})^T$. The examples show that the data X_t can be naturally or artificially multivariate.

Estimating functions are effectively generalized method of moments (also related to M-estimators, which are a particular class of estimating functions), where it is exploited that the true parameter θ_0 is the only one fulfilling $\mathrm{E}\, G(X_1, \theta_0) = 0$. In our setting, we do not require that the model is actually true but can use the model as a tool for feature extraction. In this case θ_0 is the best approximating parameter (in the above sense). This enables us to construct a change point method which will detect changes that cause a change

in the best approximating parameter (see e.g. Kirch et al. (2015) as well as Kirch and Kamgaing (2012) where this idea was exploited for offline tests).

Because $\mathrm{E}\,G(X^{(1)}, \theta_0) \neq 0$ for the observations $X^{(1)}$ after the change, we consider monitoring schemes based on means of $G(X_i, \widehat{\theta}_m)$, $i \geqslant m$. However, it is somewhat restrictive to require that the same estimating function G needs to be used, hence we allow for a different function H, which does not necessarily need to be an estimating function itself but can for example be lower-dimensional in order to increase power for certain alternatives. Furthermore, this allows to use a more precise but less robust function for estimating (e.g. if it is known that the historic data set does not contain any outliers) while using a more robust monitoring function (as there may be outliers during the observation period). See Kirch and Kamgaing (2015) for more details as well as examples for this idea.

All detector statistics are based on partial sum processes

$$S_m(l, u) := \sum_{t=m+l+1}^{m+u} H(X_t, \widehat{\theta}_m).$$

Because the monitoring function H and hence the partial sum processes are often multivariate the test decision will be based on quadratic forms $\|S\|_A^2 = S^T A S$ with a suitable matrix A resulting in the following statistics

$$
\begin{aligned}
\Gamma_1(m,k) &:= \|S_m(0,k)\|_A & \text{CUSUM,} \\
\Gamma_2(m,k) &:= \Gamma_2(m,k,h_2) := \|S_m(\lfloor kh_2 \rfloor, k)\|_A, & \text{mMOSUM,} \\
\Gamma_3(m,k) &:= \sup_{0 \leqslant i < k} \|S_m(i,k)\|_A & \text{Page-CUSUM,} \\
\Gamma_4(m,k) &:= \Gamma_4(m,k,h_4) = \|S_m(k-h_4, k)\|_A, & \text{MOSUM} \quad (2)
\end{aligned}
$$

where $\Gamma_j = \Gamma_{j,A}$ for $j = 1, \ldots, 4$ and the dependence on A is suppressed where possible for better readability. The notation $\lfloor x \rfloor$ indicates the lower Gauss bracket, i.e. the largest natural number smaller or equal than x, while $\lceil x \rceil$ indicates the upper Gauss bracket, i.e. the smallest natural number larger or equal to x.

The bandwidth $h_2 \in (0, 1)$ (fixed) is a tuning parameter determining the rate with which early observations are discarded, $h_4 = h_{4,m} \in \mathbb{N}$ is the window size in the moving sum (MOSUM) procedure.

Statistic Γ_1 has already been dealt with in detail in Kirch and Kamgaing (2015) and will not be considered here except in the simulation study and data analysis.

In order to control the asymptotic size of the detection procedure even in the presence of an infinite observation horizon, weight functions $w_j(k)$ are introduced that depend on the point of time and either the length of historic data set m (i.e. $w_j(k) = w_j(m, k)$, $j = 1, 2, 3$, for the CUSUM, mMOSUM and Page-CUSUM or the window width h_4 for the MOSUM method (i.e. $w_4(k) = w_4(h_4, k)$). As before we supress the dependence on m and h in order to obtain a unified notation.

The null hypothesis of no change will be rejected at the first point in time k when

$$w_j(k)\Gamma_j(m, k) > c_j, \quad j = 1, 2, 3, 4,$$

where c_j is a critical value which will be derived from the limit distribution under the null hypothesis. An alternative way of thinking about this is that $c_j/w_j(k)$ is a critical curve that needs to be crossed by the detector statistics $\Gamma_j(m, k)$. In Section 3.2 we will plot $w_j(k)\Gamma_j(m, k)/c_j$ for an easier visual comparison, so that all detector statistics need to cross the horizontal 1-line.

The stopping times of the corresponding sequential procedures are defined as

$$\tau_{j,m} = \min\{1 \leqslant k < N(m) : \ w_j(k)\Gamma_j(m, k) > c_j\},$$

where we set $\min \emptyset = \infty$. We distinguish between the open-end procedure with $N(m) = \infty$ and the closed-end procedure with $N(m) = mN + 1, N > 0$.

If the weight function and critical value are chosen appropriately, the type-I-error is controlled and the method has asymptotic power one under alternatives, i.e.

$$\lim_{m\to\infty} P_{H_0}(\tau_{j,m} < \infty) = \lim_{m\to\infty} P_{H_0}\left(\sup_{1\leqslant k < N(m)} w_j(k)\Gamma_j(m, k) > c_j\right) = \alpha,$$

$$\lim_{m\to\infty} P_{H_1}(\tau_{j,m} < \infty) = 1.$$

The matrix A is usually chosen such that the null limit of $\sup_{1\leqslant k < N(m)} w_j(k)\Gamma_j(m, k)$ is pivotal, e.g. $A^{-1} = \text{cov}(H(X_1, \theta_0))$ for i.i.d. data or $A^{-1} = \lim_{n\to\infty} \text{cov}\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{n} H(X_j, \theta_0)\right)$ the long-run covariance matrix in case of a time series.

1.2. Examples

The mMOSUM has, to the best of our knowledge, only been considered in the linear model in Chen and Tian (2010), where – due to an error in the proof – the asymptotic limit distribution under the null hypothesis is incorrect. The Page-CUSUM has been considered in Fremdt (2014) also for the linear model. The MOSUM statistic has been analysed in the location model in Horváth et al. (2008) and Aue et al. (2008).

Kirch and Kamgaing (2015) develop the theory for the CUSUM statistic in an equivalent setting to the one investigated in this paper. To unify the theory for all statistics we use the same assumptions in this paper to prove the corresponding limit results under the null hypothesis with the exception of the MOSUM statistic with $h_4/m \to 0$, which needs somewhat different assumptions. Therefore, the validity of the procedures based on the mMOSUM, Page-CUSUM as well as MOSUM (with $h/m \to \beta > 0$) monitoring schemes follows for all examples considered in Kirch and Kamgaing (2015). Similarly, under alternatives

we give a stricter assumption than necessary in order to unify the theory for all statistics, which can be relaxed (see Weber (2017), Chapter 4). We will now give a short overview over these examples, more details can be found in Kirch and Kamgaing (2015).

1.2.1. Linear regression model

The model is given by

$$X_t = \boldsymbol{x}_t^T \boldsymbol{\beta}_t + \epsilon_t, \ 0 \leqslant t < \infty,$$

where $\boldsymbol{\beta}_t = (\beta_{t,1}, \ldots, \beta_{t,p})^T$ are the unknown parameters and $\boldsymbol{x}_t = (1, x_{t,2}, \ldots, x_{t,p})^T$ are regressors. Moreover, the errors $\{\epsilon_t\}$ have mean zero and variance $\sigma_t^2 = \sigma_0^2$ and are independent of $\{\boldsymbol{x}_t\}$. In the literature the residuals are often supposed to be i.i.d. or uncorrelated with some moment conditions, but the minimal requirement is that they fulfill a functional central limit theorem.

As before, we consider the null hypothesis of no change ($H_0 : \boldsymbol{\beta}_t = \boldsymbol{\beta}_0$ for all t) against the alternative hypothesis of a change at $m + k^*$ ($H_A : \boldsymbol{\beta}_t = \boldsymbol{\beta}_0$ for $t \leqslant m + k^*$ but $\boldsymbol{\beta}_t = \boldsymbol{\beta}_A \neq \boldsymbol{\beta}_0$ for $t > m + k^*$). Here, k^* is the change point, and m denotes the length of the historical data set.

Both Chen and Tian (2010) for the mMOSUM as well as Fremdt (2014) for the Page-CUSUM use the least squares estimator to obtain an approximation for the unknown parameter $\boldsymbol{\beta}_0$ based on the historical data set. The corresponding estimating function is given by $G((X_t, \boldsymbol{x}_t^T)^T, \boldsymbol{\beta}) = \boldsymbol{x}_t(X_t - \boldsymbol{\beta}^T \boldsymbol{x}_t)$. Their monitoring function for parameter changes is given by $H((X_t, \boldsymbol{x}_t^T)^T, \boldsymbol{\beta}) = X_t - \boldsymbol{\beta}^T \boldsymbol{x}_t$ with $B(\theta_0) = (1, 0, \ldots, 0)^T$, i.e. their detector is based on the estimated residuals. However, this monitoring function cannot detect all alternatives so that Hušková and Koubková (2005) propose the CUSUM statistic with $H = G$ resulting in a test with asymptotic power one for all alternatives.

In order to find changes in the error variance σ_t^2, we can extend the estimating function to $\tilde{G}((X_t, \boldsymbol{x}_t^T)^T, (\boldsymbol{\beta}, \sigma^2)) = (G((X_t, \boldsymbol{x}_t^T)^T, \boldsymbol{\beta})^T, (X_t - \boldsymbol{\beta}^T \boldsymbol{x}_t)^2 - \sigma^2)^T$ and $\tilde{H}((X_t, \boldsymbol{x}_t^T)^T, (\boldsymbol{\beta}^T, \sigma^2)^T) = (X_t - \boldsymbol{\beta}^T \boldsymbol{x}_t)^2 - \sigma^2$ with $B(\theta_0) = (0, \ldots, 0, 1)^T$. The corresponding procedures have also been considered by Chen and Tian (2010) as well as Fremdt (2014) respectively.

1.2.2. Location model

This is the simplest change point setting, where

$$X_t = \mu + \Delta \mathbf{1}_{\{t > m + k^*\}} + \epsilon_t, \ \ 0 \leqslant t < \infty,$$

where k^* is the change point, $\mu \in \mathbb{R}$ is the mean prior the change, $\Delta \neq 0$, and ϵ_t is an error sequence with mean zero and variance σ^2. The multivariate mean change model is also included in the general setup but for simplicity we concentrate on the univariate model.

In fact, this model is a special case of the previous regression model with $p = 1$. Estimating and monitoring function simplify to $G(X_t, \mu) = X_t - \mu$ resulting in $\widehat{\mu} = \bar{X}_m$ and $H = G$. Horváth et al. (2008) and Aue et al. (2008) investigate this procedure for the MOSUM detector.

It is also possible to use more robust monitoring procedures e.g. based on M-estimators in this context and one can even combine a non-robust initial estimator with a robust monitoring function by choosing the monitoring function appropriately and different from the estimating function. For more details, we refer to Kirch and Kamgaing (2015), Section 6.2.

1.2.3. Non-linear models

Several monitoring procedures based on the CUSUM detectors that fit into the above framework have already been investigated in the literature including one for GARCH-sequences (Berkes et al., 2004), for nonlinear regression models with $Y_t = f(\boldsymbol{x}_t, \boldsymbol{\beta}_0) + \epsilon_t$ (Ciupera, 2013) as well as nonlinear autoregressive time series $X_t = g(X_{t-1}, \ldots, X_{t-p}) + \epsilon_t$ for some function g where a neural network approximation is used to construct the detectors (Kirch and Kamgaing, 2015). More details can be found in Kirch and Kamgaing (2015), Section 6.3.

1.2.4. Integer-valued time series

A binary autoregressive time series is given by

$$X_t | X_{t-1}, X_{t-2}, \ldots, Z_{t-1}, Z_{t-2}, \cdots \sim \mathrm{Bern}(\pi_t(\boldsymbol{\beta})), \quad g(\pi_t(\boldsymbol{\beta})) = \boldsymbol{\beta}^T \boldsymbol{Z}_{t-1},$$

where $\boldsymbol{Z}_{t-1} = (Z_{t-1}, \ldots, Z_{t-p})$ are regressors which can be purely autoregressive, purely exogenous or a combination of both. In this case a typical estimating function is based on the partial likelihood scores with a monitoring function either being of the same form or a projection onto one or more of the components of the full partial likelihood scores.

Similarly, the Poisson autoregressive model is defined by

$$X_t | X_{t-1}, \ldots, X_{t-p} \sim \mathrm{Pois}(\lambda_t), \quad \lambda_t = f_\theta(\boldsymbol{X}_{t-1}), \boldsymbol{X}_{t-1} = (X_{t-1}, \ldots, X_{t-p})^T,$$

where the estimating function (as well as monitoring function) is also typically obtained by the partial log likelihood scores. For more details we refer to Kirch and Kamgaing (2015), Sections 6.4 and 6.5.

2. Asymptotics

We first derive the null asymptotics for the respective statistics, allowing us to control the type-I-error of the procedures, before showing that the procedures are asymptotically consistent under alternatives.

2.1. Asymptotics under the null hypothesis

In this section, we give regularity conditions under which we can prove the limit distribution for the above statistics. For the CUSUM statistic this was already done in Kirch and Kamgaing (2015). It turns out that we can derive the null asymptotics for the mMOSUM, the Page-CUSUM as well as the MOSUM with $h_4/m \to \beta > 0$ under the same set of regularity conditions. In particular, the limit results carry over to all the examples given in Kirch and Kamgaing (2015) as shortly introduced in the previous section. Unlike Chen and Tian (2010) we do not use a weight function for the mMOSUM that depends on the bandwidth, although such results can also be obtained with the below methods. Their choice of weight function seems to result in a very nice limit distribution in their paper, which is however only due to an error in the proof. More precisely the second formula in the proof of Theorem 3.1 in Chen and Tian (2010) is not correct: An application of the invariance principle correctly gives (in our notation, where $\{\varepsilon_t\}$ are the innovations in a linear regression model with variance σ^2)

$$\sup_{k \geqslant 1} w(m, k - \lfloor kh_2 \rfloor) \left| \sum_{i=m+1+\lfloor kh_2 \rfloor}^{m+k} \varepsilon_i - \sigma(W_{1,m}(k) - W_{1,m}(\lfloor kh_2 \rfloor)) \right| = o_P(1)$$

for standard Wiener processes $W_{1,m}(\cdot)$. In Chen and Tian (2010) the term $W_{1,m}(k) - W_{1,m}(\lfloor kh_2 \rfloor)$ was then replaced by $W_{1,m}(k - \lfloor kh_2 \rfloor)$. However, the two processes are distributionally different, as e.g. for k_1, k_2 with $\lfloor k_1 h_2 \rfloor \geqslant k_2$ it holds due to the independent increments of a Wiener process

$$\text{cov}\,(W_{1,m}(k_1) - W_{1,m}(\lfloor k_1 h_2 \rfloor), W_{1,m}(k_2) - W_{1,m}(\lfloor k_2 h_2 \rfloor)) = 0,$$

while

$$\text{cov}\,(W_{1,m}(k_1 - \lfloor k_1 h_2 \rfloor), W_{1,m}(k_2 - \lfloor k_2 h_2 \rfloor)) \neq 0.$$

As one consequence, we can no longer substitute $k - \lfloor kh_2 \rfloor$ with s (as in the line after (20) in the proof of Theorem 3.2 in Chen and Tian (2010)), so that it is no longer beneficial to use the weight function $w(m, k - \lfloor kh_2 \rfloor)$. Therefore, we decided to return to the usual shape of the weight function – after all the bandwidth in this case is merely a way of using only more recent observations for later comparisons.

For the MOSUM statistic with $h_4/m \to 0$ we get related but somewhat different regularity conditions. Exemplary, we prove their validity for a linear regression model thus extending the existing results from the location to a linear regression model (see Section 2.2).

In the remainder of the paper, we will give some high-level regularity conditions under which we can derive the asymptotic behavior both under the null as well as under alternative hypotheses of the proposed monitoring schemes. In particular, some effort is needed to verify those regularity condition for a given change point problem and a given set of estimating/monitoring functions. Fortunately, the regularity conditions RC.1, RC.2 and RC.3 are the same that have

been used in Kirch and Kamgaing (2015) for the CUSUM monitoring scheme. In fact, these regularity conditions have already been proven to be correct by various authors in many different situations including all examples given in Section 1.2. A detailed collection of these papers is given in Kirch and Kamgaing (2015), Section 6, where also simulation results and a data analysis for various examples and the CUSUM scheme can be found. The below formulation via regularity conditions has the important advantage that a statistician who develops a new procedure in the above spirit for a new setting will immediately have all of the above monitoring schemes at his disposal (with some additional effort needed only for the MOSUM with $h_4/m \to 0$).

Regularity Condition RC.1.

(a) *The partial sum process*
$\left\{ \frac{1}{\sqrt{m}} \sum_{t=1}^{\lfloor ms \rfloor} (H(X_t, \theta_0), G(X_t, \theta_0)) : 1 \leqslant s \leqslant T \right\}$ *fulfills a functional central limit theorem for any $T > 0$:*

$$
\left\{ \frac{1}{\sqrt{m}} \sum_{t=1}^{\lfloor ms \rfloor} (H(X_t, \theta_0), G(X_t, \theta_0)) : 1 \leqslant s \leqslant T \right\}
$$
$$
\xrightarrow{\mathcal{D}} \left\{ (W_1(s), W_2(s)) : 1 \leqslant s \leqslant T \right\},
$$

where $(W_1(s), W_2(s))$ has the covariance matrix $\Sigma = \begin{pmatrix} \Sigma_1 & C \\ C^T & \Sigma_2 \end{pmatrix}$.

(b) *The following Hájek -Rényi-type inequality is fulfilled for all $0 \leqslant \alpha < \frac{1}{2}$ uniformly in m*

$$
\max_{1 \leqslant k \leqslant m} \frac{1}{m^{\frac{1}{2} - \alpha} k^\alpha} \left\| \sum_{t=m+1}^{m+k} H(X_t, \theta_0) \right\| = O_P(1).
$$

(c) *For the open-end procedure the following Hájék-Rényi-type inequality is fulfilled for any $k_m \geqslant 0$ uniformly in m*

$$
\max_{k \geqslant k_m} \frac{\sqrt{k_m}}{k} \left\| \sum_{t=m+1}^{m+k} H(X_t, \theta_0) \right\| = O_P(1).
$$

This set of regularity conditions ensures that the limit distribution is an appropriate functional of a Wiener process, where (b) is needed to control the behavior at the very beginning of the monitoring period, while (c) is needed to control the behavior at the infinite end of the monitoring period.

The regularity condition in (a) holds for example under mixing conditions on $\{X_t\}$. Obviously $H(X_t, \theta_0)$ and $G(X_t, \theta_0)$ will typically be highly correlated (often even being equal), but the actual form of C is not important (with the exception of the MOSUM procedure with fixed β and an overlap of the monitoring and historic data set – see Theorem 2.3 and Remark 2.1). In all other cases, the limit only depends on the joint convergence of $\frac{1}{\sqrt{m}} \sum_{j=1}^{m} G(X_j, \theta_0)$ and

$\left\{ \frac{1}{\sqrt{m}} \sum_{j=m+1}^{m+k} H(X_i, \theta_0) : k \geqslant 1 \right\}$, which are asymptotically independent (under the above condition) due to the independent increments of a Wiener process.

Because of the different structures for the CUSUM, mMOSUM and Page-CUSUM on the one hand, and MOSUM on the other hand, slightly different regularity conditions are needed for the latter case (if $h_4/m \to 0$). Thus, we first discuss the situation of the first three (where the results for the CUSUM can already be found in Kirch and Kamgaing (2015) and will not be repeated), before giving related but slightly different assumptions for the MOSUM statistic.

2.1.1. Modified MOSUM and Page-CUSUM

Regularity Condition RC.2.

  a) *The weight function has the following form*

$$w(m, k) = m^{-\frac{1}{2}} \widetilde{w}(m, k)$$

  with

$$\widetilde{w}(m, k) = \begin{cases} \rho\left(\frac{k}{m}\right), & k \geqslant a_m \\ 0, & k < a_m \end{cases}$$

  and $\frac{a_m}{m} \overset{m \to \infty}{\longrightarrow} 0$. In addition we need that $\rho : [0, \infty] \to \mathbb{R}^+$ is a positive and continuous function fulfilling

$$\lim_{t \to 0} t^\gamma \rho(t) < \infty \quad \text{for some} \quad 0 \leqslant \gamma < \frac{1}{2}.$$

  b) *For the open-end procedure we additionally need*

$$\limsup_{t \to \infty} t\rho(t) < \infty.$$

This set of regularity conditions is concerned with the regularity of the weight function, which ensures that we can control the type-I-error even in the open-end case. The case distinction in (a) is quite useful to allow for false alarms within the very first observations after the monitoring starts (compare with Figure 1). A standard weight function often used in the literature fulfilling these assumptions is given by

$$w(m, k) = m^{-1/2} \left(1 + \frac{k}{m}\right)^{-1} \left(\frac{k}{m+k}\right)^{-\gamma}, \qquad 0 \leqslant \gamma < \frac{1}{2}. \tag{3}$$

This corresponds to the choice $\rho(t) = t^{-\gamma} (1 + t)^{\gamma - 1}$.

Regularity Condition RC.3. *The following approximation holds under H_0, where $N(m)$ is the observation horizon and can be infinite:*

$$\sup_{1 \leqslant k < N(m)} \min \left(\frac{1}{m^{\frac{1}{2}-\gamma} k^{\gamma}}, \frac{m^{1/2}}{k} \right) \left\| \sum_{i=m+1}^{m+k} H(X_i, \widehat{\theta}_m) \right.$$
$$\left. - \left(\sum_{t=m+1}^{m+k} H(X_t, \theta_0) - \frac{k}{m} B(\theta_0) \sum_{j=1}^{m} G(X_j, \theta_0) \right) \right\| = o_P(1)$$

for some θ_0 and a suitable $B(\theta_0)$, where γ is as in RC.2.

This assumption allows to replace the partial sum process including the estimated parameter value with the appropriate partial sum process with the true or best approximating parameter θ_0. The second summand may be surprising at first but this accounts for the uncertainty from using the estimator rather than the true value. A very simple example is the location model with $H(X_t, \mu) = G(X_t, \mu) = X_t - \mu$, so that $\widehat{\mu}_m = \bar{X}_m$ and

$$\sum_{i=m+1}^{m+k} H(X_i, \widehat{\mu}_m) = \sum_{i=m+1}^{m+k} X_i - \frac{k}{m} \sum_{i=1}^{m} X_i = \sum_{i=m+1}^{m+k} H(X_i, \mu) - \frac{k}{m} \sum_{i=1}^{m} G(X_i, \mu).$$

For sufficiently smooth estimating and monitoring functions, this condition can be derived by a Taylor expansion under weak moment conditions with

$$B(\theta_0) = \mathrm{E} \, \nabla H(\mathbf{X}_0, \theta_0) \, \left(\mathrm{E} \, \nabla G(\mathbf{X}_0, \theta_0) \right)^{-1},$$

where ∇ is the gradient for a vector-valued function $F = (F_1, \ldots, F_d)^T : \mathbb{R}^d \to \mathbb{R}^d$ defined by $\nabla F = (\nabla F_1, \ldots, \nabla F_d)$ and ∇F_1 denotes the standard gradient. Details are given in Kirch and Kamgaing (2015), Section 5.

In particular, whenever H is a linear combination of $G_1, \ldots, G_d$ with $G = (G_1, \ldots, G_d)^T$, then $B(\theta_0) G(\mathbf{X}, \theta_0) = H(\mathbf{X}, \theta_0)$ and hence $\Sigma_1 = B(\theta_0) \Sigma_2 B(\theta_0)^T$, which is important in view of Theorem 2.2 below. This includes the two standard situations that the monitoring function is either given by the estimating function or by a projection onto one particular component of the estimating function (such as estimated residuals if e.g. a least squares estimator is used in a linear regression context). Most monitoring schemes developed in the literature are of this simpler type. Exceptions are the monitoring schemes based on estimated residuals in a non-linear setting proposed by Ciupera (2013) as well as the robust monitoring scheme combined with a more precise (but not robust) estimating function in Kirch and Kamgaing (2015), Section 6.2.

Under these assumptions we can derive the following null asymptotics for the mMOSUM Γ_2 and the Page-CUSUM Γ_3:

Theorem 2.1. *Let Regularity Conditions RC.1a), RC.2a) as well as RC.3 be fulfilled. Then, we get the following limit distributions (for $m \to \infty$ and fixed bandwidth $h_2 > 0$) under the null hypothesis for any symmetric positive semidefinite matrix A with the notation as in (2).*

(a) *Closed-end procedure*
   For a bounded weight function ρ or if additionally RC.1(b) holds, we get:

   (i) $\displaystyle\sup_{1\leqslant k\leqslant Nm} w(m,k)\Gamma_2(m,k,h_2)$

   $$\xrightarrow{\mathcal{D}} \sup_{0<t\leqslant N} \rho(t)\left\|W_1(1+t)-W_1(1+th_2)-t(1-h_2)W_2(1)\right\|_A,$$

   (ii) $\displaystyle\sup_{1\leqslant k\leqslant Nm} w(m,k)\Gamma_3(m,k)$

   $$\xrightarrow{\mathcal{D}} \sup_{0<t\leqslant N} \rho(t)\max_{0\leqslant s\leqslant t}\left\|W_1(1+t)-W_1(1+s)-(t-s)W_2(1)\right\|_A$$

(b) *Open-end procedure*
   If additionally regularity conditions RC.2b) and RC.1b)-c) hold, then we get:

   (i) $\displaystyle\sup_{1\leqslant k<\infty} w(m,k)\Gamma_2(m,k,h_2)$

   $$\xrightarrow{\mathcal{D}} \sup_{t>0} \rho(t)\left\|W_1(t)-W_1(th_2)-t(1-h_2)W_2(1)\right\|_A,$$

   (ii) $\displaystyle\sup_{1\leqslant k<\infty} w(m,k)\Gamma_3(m,k)$

   $$\xrightarrow{\mathcal{D}} \sup_{t>0} \rho(t)\max_{0\leqslant s\leqslant t}\left\|W_1(t)-W_1(s)-(t-s)W_2(1)\right\|_A$$

In all cases $\{W_1(t) : t \geqslant 0\}$ and $\{W_2(t) : t \geqslant 0\}$ are independent Wiener processes with covariance matrices Σ_1 and $B(\theta_0)\Sigma_2 B(\theta_0)^T$ with Σ_j, $j = 1,2$, as in Assumption RC.1a) and $B(\theta_0)$ as in RC.3.
The matrix A in $\Gamma_j = \Gamma_{j,A}$, $j = 2,3$, can be replaced by a consistent estimator.

The Wiener processes in the above theorem are different from the ones in RC.1a). In fact, the Wiener processes in the above limit are given by $\{W_1(s-1) : s \geqslant 1\}$ and $B(\theta_0)\,W_2(1)$ (where W_j are now as in RC.1a)). The covariance structure as given in the above theorem is thus obtained from the independent increment property.

As explained beneath RC.3 in many important cases it holds $\Sigma_1 = B(\theta_0)\Sigma_2 B(\theta_0)^T$. In this case the limit for a weight function as in (3) simplifies because it can then be expressed as a supremum over $[0,1]$ (rather than over an unbounded set).

Theorem 2.2. *If* $\Sigma_1 = B(\theta_0)\Sigma_2 B(\theta_0)^T$, *then it holds with a weight function as in* (3)

$$(a) \quad \sup_{t>0} \left\| \frac{W_1(1+t) - W_1(1+th_2) - t(1-h_2)W_2(1)}{(1+t)\left(\frac{t}{1+t}\right)^\gamma} \right\|_A$$

$$\overset{\mathcal{D}}{=} \sup_{0<s<1} \left\| \frac{W_1(s)}{s^\gamma} - (1-(1-h_2)s)\frac{W_1\left(\frac{h_2 s}{1-(1-h_2)s}\right)}{s^\gamma} \right\|_A,$$

$$(b) \quad \sup_{t>0} \max_{0 \leqslant s \leqslant t} \left\| \frac{W_1(1+t) - W_1(1+s) - (t-s)W_2(1)}{(1+t)\left(\frac{t}{1+t}\right)^\gamma} \right\|_A$$

$$\overset{\mathcal{D}}{=} \sup_{0<t<1} \max_{0 \leqslant s \leqslant t} \frac{1}{t^\gamma} \left\| W_1(t) - \frac{t-1}{s-1}W_1(s) \right\|_A,$$

where $\{W_1(\cdot)\}$ *and* $\{W_2(\cdot)\}$ *are as in Theorem* 2.1.

2.1.2. MOSUM

Since the MOSUM statistic has a different weight function, we need to replace Regularity Condition RC.2:

Regularity Condition RC.4. *The weight function has the following form*

$$w(h_4, k) = h_4^{-\frac{1}{2}} \rho\left(\frac{k}{h_4}\right) \tag{4}$$

where ρ *is a bounded and continuous function. Depending on the procedures additional assumptions on the behavior of* ρ *at infinity have to be made that will be specified in the theorems below.*

For the MOSUM procedure one needs to distinguish two main cases: The case where the bandwidth h_4 is of the same order as the historic data set i.e. when $h_4 = \beta m + o(m)$ for some $0 < \beta \leqslant 1$ and the situation where it is much smaller, i.e. $h_4/m \to 0$.

In the first case, we can use the same regularity conditions on the time series as before, in the latter case, somewhat different regularity conditions are needed. In particular, the assertion of Theorem 2.3 below holds true for all the examples of Section 1.2, see Section 6 in Kirch and Kamgaing (2015) for proofs of the regularity conditions.

Theorem 2.3. *Let Regularity Conditions RC.1 (a), RC.3 with $\gamma = 0$ and RC.4 hold. Then, we get the following limit theorem under the null hypothesis and for $h_4 = \beta m + o(m)$, $0 < \beta \leqslant 1$:*

a) *Closed-end procedure*
For any symmetric positive semi-definite matrix A, we get as $m \to \infty$

$$\sup_{1 \leqslant k \leqslant Nm} w(h_4, k) \Gamma_4(m, k, h_4)$$

$$\xrightarrow{\mathcal{D}} \sup_{0 < t \leqslant \frac{N}{\beta}} \rho(t) \left\| W_1 \left(\frac{1}{\beta} + t \right) - W_1 \left(\frac{1}{\beta} + t - 1 \right) - \beta W_2 \left(\frac{1}{\beta} \right) \right\|_A.$$

b) *Open-end procedure*
If additionally RC.1(c) holds and ρ fulfills RC.2 (b), then

$$\sup_{1 \leqslant k < \infty} w(h_4, k) \Gamma_4(m, k, h_4)$$

$$\xrightarrow{\mathcal{D}} \sup_{0 < t < \infty} \rho(t) \left\| W_1 \left(\frac{1}{\beta} + t \right) - W_1 \left(\frac{1}{\beta} + t - 1 \right) - \beta W_2 \left(\frac{1}{\beta} \right) \right\|_A.$$

Here, $\{(W_1(t), W_2(t)) : t \geqslant 0\}$ is a Wiener processes with covariance matrix

$$\begin{pmatrix} \Sigma_1 & CB(\theta_0)^T \\ B(\theta_0)\,C^T & B(\theta_0)\Sigma_2 B(\theta_0)^T \end{pmatrix}$$

with C, Σ_j, $j = 1, 2$, as in Assumption RC.1a) and $B(\theta_0)$ as in RC.3.
The matrix A in $\Gamma_4 = \Gamma_{4,A}$ can be replaced by a consistent estimator.

Remark 2.1. *If $\rho(t) = 0$ for $t < 1$, i.e. if the supremum is only taken over $k \geqslant h_4$, then due to the independent increment property of Wiener processes the limit simplifies to (with possibly $N = \infty$)*

$$\sup_{1 < t < Nm} \rho(t) \left\| W_1(t) - W_1(t - 1) - \beta W_2 \left(\frac{1}{\beta} \right) \right\|_A$$

for independent Wiener processes $\{W_1(\cdot)\}$ and $\{W_2(\cdot)\}$ with covariance matrices Σ_1 and $B(\theta_0)\Sigma_2 B(\theta_0)^T$ respectively. Under RC.1(b) and if ρ fulfills the assumption in RC.2 (a) for $t \to 0$ one can also use the lower bound $\max(m + 1, m + k - h_4 + 1)$ resulting also in a non-overlapping MOSUM.

The condition RC.2 (b) is too strict for some weight functions that have been proposed in the literature in the case of $h_4/m \to 0$. If one wants to relax them more moments of the process are required, in order to get a proper limit distribution for these weight functions (see Horváth et al. (2008) for an extensive discussion).

In order to make this more precise, we need the following stronger assumptions:

Regularity Condition RC. 5. *Let $\{X_t\}$ (under H_0) be stationary and assume there exists (possibly after changing the probability space) a Wiener process $\{W_1(t), 0 \leqslant t < \infty\}$ with covariance matrix Σ_1 such that, as $k \to \infty$*

$$\sum_{t=1}^{k} H(X_t, \theta_0) - W_1(k) = O\left(k^{\frac{1}{\nu}} \right) \quad a.s. \quad for \ some \ \nu > 2. \tag{5}$$

The value of ν usually corresponds to the number of moments of $H(X_t, \theta_0)$. The larger it is, the closer the partial sum process is already to a Wiener process. We will see that this allows us to use boundary functions that grow to zero slower at infinity. Invariance principles as above have been proven for many different time series, e.g. they hold for mixing time series $\{X_t\}$ with sufficiently fast mixing rate (see also Kirch and Kamgaing (2015)).

The following regularity condition is somewhat different from the one in RC.3 and has to the best of our knowledge only been verified for mean changes. In Section 2.2 we prove exemplary its validity for the linear regression case with least squares estimation.

Regularity Condition RC.6. *The following approximation holds under H_0, where $N(m)$ is the observation horizon and can be infinite:*

$$\sup_{1 \leqslant k < N(m)+1} \min(h_4^{-1/2}, h_4^{1/\nu-1/2} k^{-1/\nu}) \left\| \sum_{i=m+k-h_4+1}^{m+k} H(X_t, \widehat{\theta}_m) \right.$$
$$\left. - \left(\sum_{i=m+k-h_4+1}^{m+k} H(X_t, \theta_0) - \frac{h_4}{m} B(\theta_0) \sum_{j=1}^{m} G(X_j, \theta_0) \right) \right\| = o_P(1)$$

with ν as in RC.5.

We can now also state the assertion for $h_4/m \to 0$.

Theorem 2.4. *Let Regularity Conditions RC.4, RC.5 and RC.6 hold. Let the weight function ρ fulfill $\limsup_{t \to \infty} t^{1/\nu} \rho(t) < \infty$, ν as in RC.5. Then, we get the following limit theorem under the null hypothesis and for $h_4/m \to 0$ for the closed-end (with $N(m) = Nm$) as well as open-end procedure with $N(m) = \infty$:*
For any symmetric positive semi-definite matrix A, we get

$$\sup_{1 \leqslant k < N(m)+1} w(h_4, k) \, \Gamma_4(m, k, h_4) \xrightarrow{\mathcal{D}} \sup_{0 < t < \infty} \rho(t) \, \|W(t+1) - W(t)\|_A \,,$$

where $\{W(t) : t \geqslant 0\}$ is a Wiener process with covariance matrix Σ_1. The matrix A in $\Gamma_4 = \Gamma_{4,A}$ can be substituted by a consistent estimator.

Remark 2.2. *If ρ fulfills RC.2 (b) we can replace RC.5 by RC.1(a) and (c) and need RC.6 with $\nu = 1$ in addition to stationarity under H_0 with an analogous proof to the proof of Theorem 2.1.*

Because the above theorem still excludes some weight functions that have been proposed in the literature such as $\rho(t) = \max(1, \log(1+t))^{-1/2}$ even in the closed-end procedure, Horváth et al. (2008) suggested to use an h_4-closed-end procedure in the sense that the observation horizon ends after Nh_4 observations for some finite N. In this case, we get the following limit distribution:

Corollary 2.1. *Let RC.1 (a), RC.4 as well as RC.6 (with any ν) hold. Then we get under the null hypothesis for any symmetric positive definite matrix A*

$$\sup_{1 \leqslant k \leqslant Nh_4} w(h, k) \, \Gamma_4(m, k, h_4) \xrightarrow{\mathcal{D}} \sup_{0 < t < N} \rho(t) \, \|W(t+1) - W(t)\|_A \,,$$

where $\{W(t) : t \geqslant 0\}$ is a Wiener process with covariance matrix Σ_1. The matrix A in $\Gamma_4 = \Gamma_{4,A}$ can be substituted by a consistent estimator.

2.2. Linear regression model

In this section, we will prove the above regularity condition RC.6 for the MO-SUM procedure with $h_4/m \to 0$ for the linear regression model as in Section 1.2.1, showing that this condition can principally be extended to other situations as well. Regularity condition RC.5 on the stochastic process have already been used by Kirch and Kamgaing (2015) and shown to be valid in a large variety of situations.

Theorem 2.5. *Let $\frac{h_4}{m} \to 0$ and the errors $\{\varepsilon_i\}$ be i.i.d. and independent of the regressors $\{\boldsymbol{x}_i\}$. Furthermore, let $\{\boldsymbol{x}_t\}$ be stationary and satisfy*

$$\frac{1}{n} \sum_{t=1}^{n} \mathbf{x}_t \mathbf{x}_t^T - \mathbf{C} = O(n^{-\tau}) \quad a.s., \tag{6}$$

for a positive definite matrix $\mathbf{C}$ and some $\tau > 0$. Then, we get for the functions G and H as in Section 1.2.1:

(a) RC.6 holds for the closed-end procedure with time horizon Nh_4.
(b) If additionally the regressors $\{\boldsymbol{x}_i\}$ fulfill a strong invariance principle

$$\sum_{i=1}^{k} (\boldsymbol{x}_t - \boldsymbol{c}_1) - \boldsymbol{W}(k) = O(k^{\frac{1}{\nu}}) \quad a.s., \quad \nu > 2,$$

where $\boldsymbol{c_1}$ the first column of the matrix $\boldsymbol{C}$ and $\boldsymbol{W}$ is a Wiener process with some covariance matrix $\boldsymbol{\Sigma}$. Then RC.6 also holds for the closed-end procedure with time horizon Nm as well as the open-end procedure.

2.3. Consistency under alternatives

In this section, we will show that under mild conditions all monitoring schemes will stop in finite time with probability approaching one as $m \to \infty$ if a change occurs. This means that the corresponding testing procedures have asymptotic power one.

Under the below regularity conditions Kirch and Kamgaing (2015) have shown this property for the CUSUM statistic. The same holds true for the other three statistics. For notational ease, we let

$$X_t = \begin{cases} Z_t, & t \leqslant m + k^*, \\ Y_t, & t > m + k^*. \end{cases}$$

Regularity Condition RC.7.

(a) *The time series before the change fulfills*

$$\sum_{t=m+1}^{m+k^*} H(X_t, \widehat{\theta}_m) = o_P(k^*).$$

(b) *$\{Y_t\}$ is stationary and independent of $\widehat{\theta}_m$ and it holds as $l \to \infty$*

$$\frac{1}{l} \left\| \sum_{j=1}^{l} \left(H(Y_j, \widehat{\theta}_m) - E_H \right) \right\| = o_P(1)$$

for some E_H.

(c) *It holds $A^{1/2} E_H \neq 0$.*

(d) *It holds $k^*/m \leqslant \lambda$ (for statistics Γ_j with $j = 1, 2, 3$) and $k^*/h \leqslant \lambda$ for the MOSUM statistic Γ_4 for some $\lambda > 0$. Furthermore, there exists a ball $U(x_0)$ around x_0 with $x_0 > \lambda$ and $\rho(x) \geqslant c > 0$ for $x \in U(x_0)$ (where we set $\rho(x) = 0$ if xm (resp. xh in the MOSUM case) is larger than the observation horizon).*

(e) *In the open-end procedures with $j = 1, 2, 3$ for an arbitrary late change k^* it holds $\liminf_{x \to \infty} x \rho(x) > 0$.*

Regularity Condition RC.7 (a) follows from the results of the previous section, if $\{Z_t\}$ fulfills the assumptions under the null hypothesis. Condition (b) is stronger then necessary but unifies the treatment for all schemes. If this assumption is not fulfilled it can be replaced by a corresponding condition, where the lower and upper limit of the sum depend on the monitoring scheme at hand (see Kirch and Kamgaing (2015), Section 4 in Weber (2017) or proofs below for details). In fact, we do not need the independence of the time series after the change $\{Y_t\}$ from the estimator $\widehat{\theta}_m$ (or the time series before the change $\{Z_t\}$) nor do we need the stationarity of $\{Y_t\}$ allowing e.g. for starting values from $\{Z_t\}$ in an autoregressive setup. This is discussed in detail in Section 5.2 in Kirch and Kamgaing (2015), where among other things this is shown for mixing sequences in combination to some smoothness assumptions on H. Conditions (d) and (e) are conditions on the weight function in connection with the location of the change but are rather mild.

Condition (c) is the key condition in the sense that it tells us what changes can be detected with a given monitoring function H and a given matrix A.

Theorem 2.6. *Let the alternative hypothesis hold in addition to regularity conditions RC.7 (a)-(c). For the closed-end-procedure let assumption (d) be fulfilled, for the open-end procedure either (d) or for all statistics except the MOSUM statistic (e). Then it holds (as $m \to \infty$)*

$$\sup_{1 \leqslant k < N(m)} w_j(k) \Gamma_j(m, k) \xrightarrow{P} \infty.$$

for all $j = 1 - 4$ for the closed-end procedures with $N(m) = Nm + 1$ as well as the open-end procedures with $N(m) = \infty$ (with the exception of the MO-SUM statistic). The matrix A in $\Gamma_j = \Gamma_{j,A}$ can be substituted by a consistent estimator.

Remark 2.3. *Because the sums in the MOSUM statistic have fixed length, we can not allow for an abritarily late change. However, if $k^* = o(h^{1+\nu/2})$ and $\liminf_{x \to \infty} x^{1/\nu} \rho(x) > 0$, then it is detectable for any MOSUM procedure with observation horizon larger than $k^* + h$.*

3. Simulation study and data analysis

In this section, we aim at illustrating the differences between the discussed monitoring schemes by some simulations but also by a comparative application to two real data sets. For this purpose we will use the well-known examples of a mean change and a linear regression model because even in this case, no comparative simulation study involving all four monitoring schemes exists in the literature (to the best of our knowledge). We refer the reader who is interested in the performance of this methodology in the different statistical situations that have been discussed in Section 1.2 to Kirch and Kamgaing (2015), Section 6, where simulations results as well as data analyses for the CUSUM statistic and many examples beyond linear regression are given in detail.

3.1. Simulation study

In this section we compare all four monitoring schemes in terms of their size, power and run length i.e. the time until an alarm is given. Because a higher empirical size leads to a better power and shorter run-length, we can only compare the power in a meaningful way when we fix the size. So for a comparison we use the size-corrected power and size-corrected run-length i.e. the power respectively run-length corresponding to the true (not nominal) size α. Furthermore, we give a kind of density estimator of the run-length which we scale so that it integrates to the size-corrected power (rather than one). One can think of this as a probability distribution with a continuous part as given by our density estimation and a discrete part at infinity (indicating that the procedure never rejected).

All of the empirical results are based on a training period of length $m = 100$, a monitoring period of 200 and 1000, 2500 repetitions and standard normal errors. For the CUSUM, Page-CUSUM and mMOSUM we use the boundary function (3) and for the MOSUM we use

$$w(h,k) = \left(2 \max\left\{1, \log\left(1 + \frac{k}{h}\right)\right\}\right)^{-\frac{1}{2}} \tag{7}$$

which has been proposed by Horváth et al. (2008) and Aue et al. (2008) (with the limit as in Corollary 2.1). For the MOSUM statistic N is chosen so that we

Monitoring per.	$N = 2$						$N = 10$	
	$\gamma = 0$		$\gamma = 0.25$		$\gamma = 0.45$		$\gamma = 0$	
CUSUM	1.12	(2.96)	2.32	(5.92)	3.36	(5.96)	3.4	(7.68)
Page-CUSUM	0.96	(2.76)	1.8	(5.32)	2.16	(3.84)	3.2	(7.72)
mMOSUM $a_m = 1$								
$h_2 = 0.1$	0.88	(3)	1.92	(4.48)	3.04	(6.88)		
$h_2 = 0.4$	1.76	(3.68)	2.88	(6.08)	4.08	(8.84)		
$h_2 = 0.9$	4.4	(7.76)	8.76	(14.24)	38.04	(44.8)		
mMOSUM $a_m = \sqrt{m} + 1$								
$h_2 = 0.1$	0.88	(3)	1.92	(4.48)	2.32	(5.08)	2.44	(5.72)
$h_2 = 0.4$	1.76	(3.68)	2.88	(6.08)	2.2	(5.04)	3.96	(8.92)
$h_2 = 0.9$	4.4	(7.76)	4.76	(8.84)	3.6	(6.36)	4.64	(9.36)
MOSUM	h_4							
	10		**20**				**20**	
	4.24	(8.4)	5.04		(10.56)		5.52	(11.16)

TABLE 1
Empirical size (in %) for a nominal level of 5 (10) %, $m = 100$, where the asymptotic critical values for the open-end procedure are used with the exception of the MOSUM statistic.

match the monitoring length of 200 resp. 1000 (i.e. $Nh = 200$ resp. $Nh = 1000$), while for the other three statistics the critical value for the open-end procedure is used).

Because we are mainly interested in a comparison of all four monitoring schemes, we use the location model $X_t = \mu + \Delta\mathbf{1}_{\{t>k^*\}} + \epsilon_t$ with $\mu = 0$ and $\Delta \neq 0$. The procedures are based on $G(X_t, \mu) = H(X_t, \mu) = X_t - \mu$. We use the true error variance $\sigma^2 = 1$.

Table 1 reports the empirical sizes for the different procedures. The MOSUM procedure holds the size quite nicely. All other procedures are conservative in particular for smaller values of γ with the exception of the mMOSUM with a large bandwidth $h_2 = 0.9$ and $\gamma \neq 0$. The problem in this case is that for the first 10 observations the detector sum only depends on one observation, namely X_{m+k}, which is quite volatile (and asymptotically negligble hence not taken care of by the asymptotic considerations). For $\gamma = 0$ this is not a problem, but other values of γ put more weight on those early detector values, resulting in too many false positives right after monitoring starts. This can clearly be seen by the false positives at the very beginning of the monitoring sequence as shown (circled in) in Figure 2 (c) and (d)). A simple solution is to wait for a_m e.g. $a_m = \sqrt{m}+1$ observations before starting the monitoring (which is permitted by Assumption RC.2). This solves the problem according to our empirical results given in the same table. Furthermore, the empirical size of the

		$N = 2$			$N = 10$
		γ			
	k^*	**0**	**0.25**	**0.45**	**0**
CUSUM	**0.1 m**	95.84	95.24	92.4	**99.76**
	0.4 m	86.36	84.48	76.24	99.6
	m	45.32	42.76	31.16	98.88
	1.5 m	13.24	11.76	8.68	98.16
Page-CUSUM	**0.1 m**	97.08	96.64	94.8	**99.76**
	0.4 m	92.16	90.76	85.36	99.64
	m	55.36	50.32	39.76	99.4
	1.5 m	14.96	12.52	9.72	98.92
mMOSUM					
$h_2 = 0.1$	**0.1 m**	**97.32**	96.88	93.76	
	0.4 m	91.92	89.72	82.08	
	m	51.48	46.28	34.4	
	1.5 m	15.04	12.28	9.24	
$h_2 = 0.4$	**0.1 m**	96	95.28	87.64	
	0.4 m	**94.84**	93.16	82.36	
	m	**74**	67.08	46.68	
	1.5 m	21	16.28	9.12	
$h_2 = 0.9$	**0.1 m**	69.52	50.36	5.04	
	0.4 m	68.8	47.08	5.04	
	m	57.6	32.64	5.04	
	1.5 m	**37.36**	18.48	5.04	
mMOSUM with $a_m = \sqrt{m}+1 = 11$					
$h_2 = 0.1$	**0.1 m**	**97.32**	96.88	95.16	99.48
	0.4 m	91.92	89.72	84.2	99.4
	m	51.48	46.24	37.08	97.12
	1.5 m	15	12.28	9.52	91.36
$h_2 = 0.4$	**0.1 m**	96	95.28	91.44	99.72
	0.4 m	**94.84**	93.12	86.6	**99.72**
	m	**74**	67.04	52.6	**99.76**
	1.5 m	21	16.24	10.6	**99.64**
$h_2 = 0.9$	**0.1 m**	69.52	62.68	38.28	93.44
	0.4 m	68.8	59.2	28.76	93.36
	m	57.56	42.48	14.92	92.64
	1.5 m	37.32	24.4	8.32	91.54
MOSUM		h_4			
		10		**20**	**20**
	0.1 m	45.92		75.68	79.16
	0.4 m	24.12		56.96	63.88
	m	10.4		30	44.64
	1.5 m	6.68		15	36.56

TABLE 2

Size-adjusted power (in %) for the size 5%
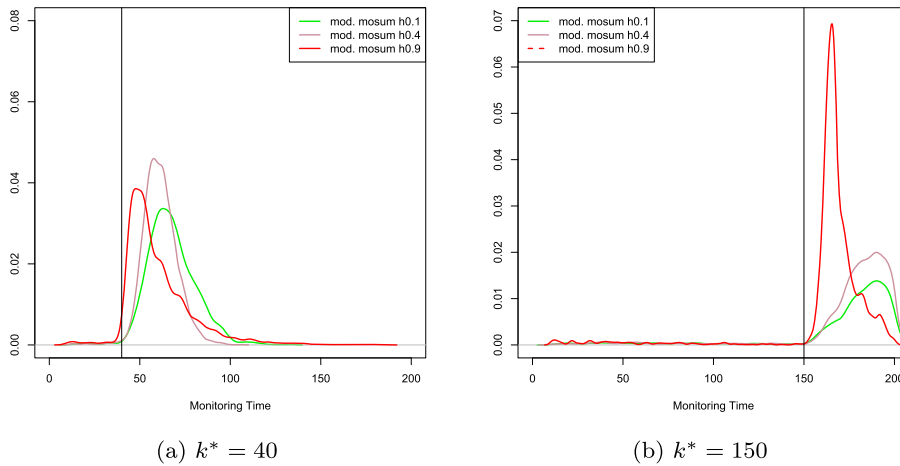
(a) $k^* = 40$ (b) $k^* = 150$

Fig 1: Run-length-plot ($\Delta = 1, \gamma = 0.25$) when monitoring only starts after $a_m = \sqrt{m}+1 = 11$ observations, level at 5%

MOSUM statistics are close to their nominal size. The other procedures use the critical values for the open-end procedure so they are in particular conservative for the smaller monitoring period. Furthermore, the asymptotic critical value is based on a supremum while the actual critical values (for normal errors) are based on a finite maximum, which shows that even with correct monitoring length the asymptotic critical values are conservative. Similarly, the plot of the run lengths in Figure 2 clearly shows a tendency of the mMOSUM with $\gamma = 0.25$ and $h_2 = 0.9$ to raise a significant number of false alarms at the very beginning of the monitoring – an effect that disappears if one starts monitoring only with an delay of a_m observations (see Figure 1). From these plots, one can clearly see that the MOSUM procedures also have some trouble with early (and in contrast to the mMOSUM not only very early) false alarms.

While for a monitoring length of $m = 100$ in this simple mean change situation, the asymptotic approximation yields reasonable results, for smaller historic lengths, more dependent data or more complicated situations, the asymptotic distribution may not yet provide good enough approximations in practise. In these cases, bootstrap methods can be helpful but have to be tailorsuited to the particular situation at hand. Kirch (2008) discuss several sequential bootstrap schemes for the mean change problem with i.i.d. errors, while Hušková and Kirch (2012) discuss the linear regression case. Both papers prove validity of the proposed method for the CUSUM statistics.

Table 2 reports the size-corrected power for all procedures. First of all, somewhat surprisingly the use of $\gamma = 0$ is always best in all situations possibly due to the fact that the other choices of γ only yield an advantage for an almost immediate change point. This can best be seen by noting that all detectors reject as soon as they are larger than the corresponding critical curve $c_j/w_j(k; \gamma)$. For a

(a) $k^* = 10, \Delta = 0.5, \gamma = 0$

(b) $k^* = 150, \Delta = 0.5, \gamma = 0$

(c) $k^* = 40, \Delta = 1, \gamma = 0.25$

(d) $k^* = 150, \Delta = 1, \gamma = 0.25$
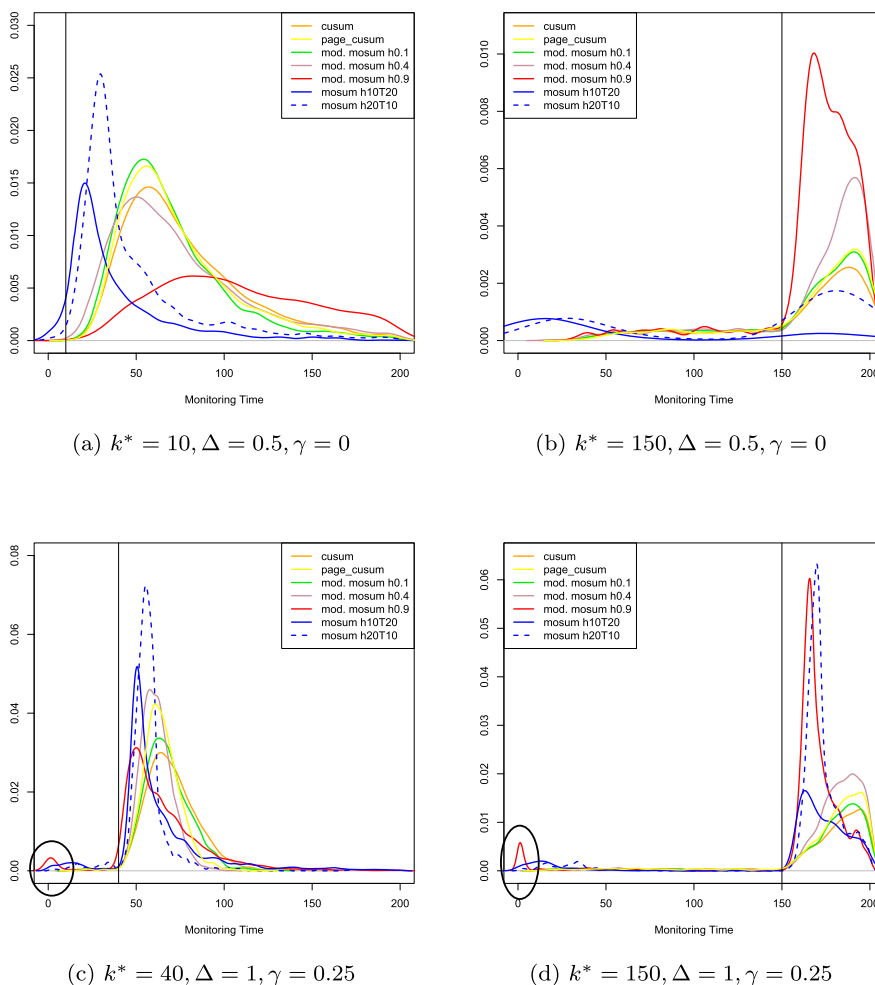
Fig 2: Run-length-plot (level at 5%)

given detector j, comparing the critical curves for different values of γ one sees that at the very beginning the curves for $\gamma \neq 0$ are below the ones for $\gamma = 0$ (which also explains the problem with the false positives at the very beginning in these cases), but very early they cross and the one with $\gamma = 0$ is beneath the others. Because of this clear superiority we only report results for $\gamma = 0$ for the longer monitoring period.

The mMOSUM detector with $h_4 = 0.4$ and somewhat less so the Page-CUSUM perform very well in all situations and are only outperformed by an mMOSUM with a bandwidth h_2 chosen according to whether an early, medium or late change occurs. To elaborate: For later changes the CUSUM detector contains more null observations than the mMOSUM detector (that disregards

the first $\lfloor h_2 k \rfloor$ observations). Therefore, more contaminated observations need to be included in case of the CUSUM before they dominate the detector enough to become significant. This fact was precisely the motivation for introducing the mMOSUM detector as well as Page-CUSUM detectors. In a sense, the bandwidth h_2 regulates how many observations (as a percentage of the current position) are discarded. Consequently, the 'best' choice of h_2 depends on the (unknown) position of the change point, i.e. the later the change is expected the larger h_2 should be chosen. In fact, this is confirmed by the simulation results in Table 2, where the best power for the different change scenarios is given in bold. At the same time, one can see that even in cases where they do not provide the detection rate, the mMOSUM with $h = 0.4$ (and somewhat less so the Page-CUSUM) do give satisfactory results, so that those are a good compromise if the user is uncertain about when changes are most likely to occur).

This is also confirmed by the plots of the run-lengths 2 as well as 1, which are rescaled so that the area under the curve integrates to the detection rate. Furthermore, these plots clearly show that the two MOSUM procedures have a quicker detection time than the other statistics, but at the cost of a much lower detection rate (and a somewhat higher false detection rate in particular at the early stages of the monitoring). While a smaller choice of h_4 gives the fast detection rate, for this choice the loss of detection power is dramatic. This effect also occurs in the a-posteriori usage of MOSUM procedures, which are mainly useful in the context of multiple change point estimation and much less so for testing (confer also Eichinger and Kirch (2018)).

Furthermore, the MOSUM procedure has the lowest power and more importantly, the power stays well away from one even if a longer monitoring period is used, where all other procedures have an empirical (size-corrected) power of almost 1 meaning they may take a while but eventually detect all change points. The MOSUM procedure on the other hand misses a significant amount of change points but those that are detected are detected quite quickly.

3.2. *Data analysis*

In this section, we analyze two data sets that have already been discussed in the literature to compare the performance of the four different monitoring schemes. One data examples shows a mean change and thus illustrates the performance in the location model. The second data set involves a linear regression model with a change in the error variance but no change in the regression coefficient allowing us to get an impression about both the behavior under the null as well as under alternatives.

We first analyse the Boston Homicide data set contained in the R-package strucchange (see Zeileis et al. (2002)) containing the monthly number of youth homicides in Boston. In early 1995 a policing initiative – the Boston-Gun-Project was started in order to lower the youth homicides. The so-called 'Operation Ceasefire' began in the late spring of 1996. Zeileis (2006) analyzed this data set in an a-posteriori change point setting not too different from ours showing that

indeed a change occurred around that time. While the data is count data that could well be modelled by a Poisson model, we will not make use of this fact in this analysis, although the theory allows for this (see also Section 6.5 in Kirch and Kamgaing (2015) or Section 10.4 in Kirch and Kamgaing (2016)). Instead we simply use it as input sequence for the location model from Section 1.2.2 based on the least-squares estimating and monitoring functions that have already been used in the previous section. The asymptotic theory derived in this paper allows to use discrete errors in that manner with the drawback that the corresponding procedure is not optimal in this situation but the advantage of not having to specify and justify a particular count time series model. In the same spirit, it illustrates the applicability of the proposed methods to non-Gaussian data.

Figure 3 gives the time series as well as the corresponding detectors $\Gamma_j(m,k)/(c_j w_j(k))$ (for $\gamma = 0$). This normalization is chosen for easier visual comparison of the different procedure as in all cases the null hypothesis is rejected as soon as the vertical 1-line is crossed. The change should occur around 50 in the time series and we consider historic data lengths of 24 (late change), 36 (medium late change) and 48 (early change).

Only the mMOSUM (with $h_2 = 0.4$) detects the change point within the monitoring period in all cases where the detection time is quicker the earlier the change occurs within the monitoring period. The MOSUM ($h_4 = 7$), CUSUM and Page-CUSUM detect the change point only within the monitoring period for $m = 48$), where in this case the MOSUM has the quickest detection rate (only merely reaching significance).

As a second example we apply our methodology to a data set from finance, where the effective exchange rate regime is often analysed by a linear regression model on other currencies. This yields valuable insight into whether the currency is allowed to fluctuate freely by market forces or whether it is to all or some extend fixed to some other currency. In our case, we consider the daily log returns of the Chinese Yuan Renminbi (CNY) and use a basket of western currencies as regressors. This is motivated by the fact that the CNY used to be fixed to the US-Dollar but has since been announced to allow to move more freely. The same data set has been used by Zeileis et al. (2010), where also much more information on the model, the motivation behind it as well as the data set can be found.

More precisely, we apply a linear regression setup to the daily log returns of the Chinese Yuan Renminbi (CNY) from July 26nd, 2005 to July 31, 2009, where we use the daily log returns of the following currencies as regressors: USD (US Dollar), JPY (Japanese yen), EUR (Euro), GBP (British pound). In the analysis of the data set in Zeileis et al. (2010) no change in the regression coefficients can be found, but a change in the residual variance has been found. For this reason, we will use the procedures based on the functions G and H from Section 1.2.1 to find changes in the regression coefficients (or actually the mean of the residuals) as well as $\widetilde{G}$ and $\widetilde{H}$ from Section 1.2.1 to find variance changes in the residuals. The results for a monitoring period of $m = 40$ can be found in Figure 4. Subfigure (a) shows the estimated residuals where the regression

(a) Time Series

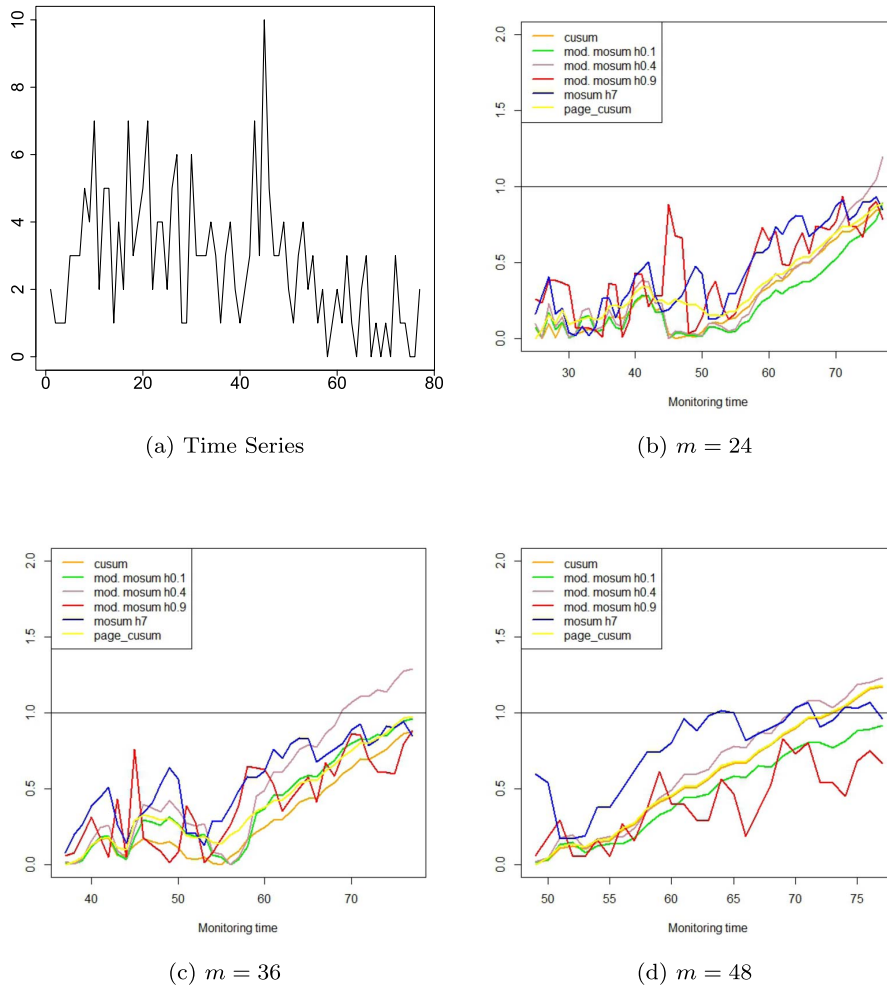(b) $m = 24$

(c) $m = 36$

(d) $m = 48$

Fig 3: Actual time series and detector statistics for Boston Homicide Data

coefficient is estimated from the historic data set. The vertical line indicates the point in time when monitoring starts. A visual inspection shows no visible mean shift but several possible changes in variance. Indeed, both assumptions are confirmed by the monitoring, with no monitoring scheme detecting a shift in the residuals (Subfigure (b)) but all of them detecting a variance change (Subfigure (a)). There are some interesting observations from the latter example: The CUSUM, Page-CUSUM and mMOSUM ($h = 0.4$) give an alarm around 90, where the increase starts around 70. This coincides with a period of smaller variability between 70 and 100 as can be seen in (a). The mMOSUM with $h = 0.9$ gives an alarm around 120, which corresponds to a series of some large variability shortly before 120. The other statistics are not influenced by this

(a) Estimated residuals

(b) Monitoring the residual mean
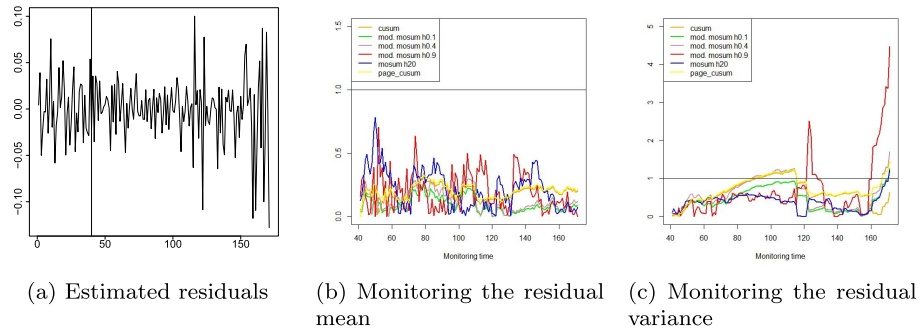
(c) Monitoring the residual variance

Fig 4: Estimated residuals and detector statistics for exchange rate data, $m = 40$ (indicated by the vertical line in (a))

because this is balanced out by the relatively small variability from before. This is also the reason why the three previously significant statistics drop under the significance line again at that point. Around 160 all monitoring schemes start to increase again and all but the CUSUM become eventually significant during the observation period. This is due to the large variability at the end of the monitoring period, where the mMOSUM ($h = 0.9$) is fastest, followed by the mMOSUM ($h = 0.4$), the Page-CUSUM, the mMOSUM ($h = 0.1$) and the MOSUM ($h = 20$). Only the CUSUM is not yet significant because all residuals from 41 to the present are used so that it takes much longer for a late change to be detected.

Both examples confirm the balanced behavior of the mMOSUM with $h = 0.4$ and to a somewhat lesser degree the Page-CUSUM detector making them preferable in many situations.

4. Proofs

Because the proofs are in parts similar, we do not provide all details of all proofs, but restrict our attention to the key steps. Missing details can be found in Weber (2017), Part I.

Proof of Theorem 2.1. By RC.2 and RC.3 it holds

$$
\sup_{1 \leqslant k < N(m)} w(m,k) \left\| \sum_{i=m+\lfloor kh_2 \rfloor+1}^{m+k} H(X_t, \widehat{\theta}_m) \right.
$$

$$
\left. - \left(\sum_{j=m+\lfloor kh_2 \rfloor+1}^{m+k} H(X_t, \theta_0) - \frac{k - \lfloor kh_2 \rfloor}{m} B(\theta_0) \sum_{j=1}^{m} G(X_j, \theta_0) \right) \right\|
$$

$$\leqslant 2\max\left(\sup_{1\leqslant k\leqslant m}\left(\frac{k}{m}\right)^{\gamma}\rho\left(\frac{k}{m}\right),\ \sup_{k\geqslant m}\frac{k}{m}\rho\left(\frac{k}{m}\right)\right)$$

$$\times\sup_{1\leqslant k<N(m)}\min\left(\frac{1}{m^{\frac{1}{2}-\gamma}k^{\gamma}},\frac{m^{1/2}}{k}\right)\left\|\sum_{i=m+1}^{m+k}H(X_t,\widehat{\theta}_m)\right.$$

$$\left.-\left(\sum_{j=m+1}^{m+k}H(X_t,\theta_0)-\frac{k}{m}B(\theta_0)\sum_{j=1}^{m}G(X_j,\theta_0)\right)\right\|=o_P(1)$$

$$(8)$$

Consequently, it is sufficient to consider the limit distribution of

$$\sup_{1\leqslant k<N(m)}w(m,k)\left\|\sum_{j=m+\lfloor kh_2\rfloor+1}^{m+k}H(X_t,\theta_0)-\frac{k-\lfloor kh_2\rfloor}{m}B(\theta_0)\sum_{j=1}^{m}G(X_j,\theta_0)\right\|.$$

With the functional limit theorem in RC.1a)i) it holds for any $\tau,N>0$

$$\sup_{\lfloor\tau m\rfloor\leqslant k\leqslant Nm}w(m,k)\left\|\sum_{j=m+\lfloor kh_2\rfloor+1}^{m+k}H(X_j,\theta_0)-\frac{k-\lfloor kh_2\rfloor}{m}B(\theta_0)\sum_{j=1}^{m}G(X_j,\theta_0)\right\|_A$$

$$=\sup_{\lfloor\tau m\rfloor\leqslant k\leqslant Nm}\sup_{\frac{k}{m}\leqslant t<\frac{k+1}{m}}\rho\left(\frac{\lfloor mt\rfloor}{m}\right)\left\|\frac{1}{\sqrt{m}}\sum_{j=m+\lfloor\lfloor mt\rfloor h_2\rfloor+1}^{m+\lfloor mt\rfloor}H(X_j,\theta_0)\right.$$

$$\left.-\frac{\lfloor mt\rfloor-\lfloor\lfloor mt\rfloor h_2\rfloor}{m}\frac{1}{\sqrt{m}}B(\theta_0)\sum_{j=1}^{m}G(X_j,\theta_0)\right\|_A$$

$$\xrightarrow{D}\sup_{\tau\leqslant t\leqslant N}\rho(t)\left\|W_1(1+t)-W_1(1+th_2)-(t-th_2)W_2(1)\right\|_A$$

For a more general weight function ρ, first note that by RC.1 (b) it holds for all $0<\alpha<\frac{1}{2}$

$$\max_{1\leqslant k\leqslant m}\frac{1}{m^{\frac{1}{2}-\alpha}k^{\alpha}}\left\|\sum_{t=m+\lfloor kh_2\rfloor+1}^{m+k}H(X_t,\theta_0)\right\|$$

$$\leqslant\max_{1\leqslant k\leqslant m}\frac{1}{m^{\frac{1}{2}-\alpha}k^{\alpha}}\left\|\sum_{t=m+1}^{m+k}H(X_t,\theta_0)\right\|$$

$$+\max_{1\leqslant k\leqslant m}\frac{h_2^{\alpha}}{m^{\frac{1}{2}-\alpha}(kh_2)^{\alpha}}\left\|\sum_{t=m+1}^{m+\lfloor kh_2\rfloor}H(X_t,\theta_0)\right\|=O_P(1)$$

From this we conclude for $\gamma < \alpha < \frac{1}{2}$

$$\sup_{1 \leqslant k < \tau m} w(m,k) \left\| \sum_{j=m+\lfloor kh_2 \rfloor + 1}^{m+k} H(X_j, \theta_0) - \frac{k - \lfloor kh_2 \rfloor}{m} B(\theta_0) \sum_{j=1}^{m} G(X_j, \theta_0) \right\|_A$$

$$\leqslant \sup_{a_m \leqslant k < \tau m} \frac{1}{\sqrt{m}} \rho\left(\frac{k}{m}\right) \left\| \sum_{j=m+\lfloor kh_2 \rfloor + 1}^{m+k} H(X_j, \theta_0) \right\|$$

$$+ \sup_{a_m \leqslant k < \tau m} \frac{k - \lfloor kh_2 \rfloor}{m^{3/2}} \rho\left(\frac{k}{m}\right) \left\| B(\theta_0) \sum_{j=1}^{m} G(X_j, \theta_0) \right\|$$

$$\leqslant \sup_{0 < t < \tau} t^\alpha \rho(t) \sup_{1 \leqslant k \leqslant \tau m} \frac{1}{m^{1/2 - \alpha} k^\alpha} \left\| \sum_{j=m+\lfloor kh_2 \rfloor + 1}^{m+k} H(X_j, \theta_0) \right\|$$

$$+ O_P(1) \sup_{0 < t < \tau} t \rho(t) = o_P(1) \tag{9}$$

as $\tau \to 0$ (uniformly in m), where in the last line also RC.1 (a) was used. This completes the proof of (a)(i). For (b) note first that by RC.1(c) it holds

$$\sup_{k \geqslant k_m} \frac{\sqrt{k_m}}{k} \left\| \sum_{t=m+1}^{m+\lfloor kh_2 \rfloor} H(X_t, \theta_0) \right\| \leqslant \sup_{k \geqslant k_m} \frac{h_2 \sqrt{k_m}}{\lfloor kh_2 \rfloor} \left\| \sum_{t=m+1}^{m+\lfloor kh_2 \rfloor} H(X_t, \theta_0) \right\|$$

$$\leqslant \sqrt{h_2 \frac{k_m h_2}{\lfloor k_m h_2 \rfloor}} \sup_{l \geqslant \lfloor k_m h_2 \rfloor} \frac{\sqrt{\lfloor k_m h_2 \rfloor}}{l} \left\| \sum_{t=m+1}^{m+l} H(X_t, \theta_0) \right\| = O_P(1).$$

From this, RC.1(c) and RC.2(b) we conclude

$$\sup_{k \geqslant Tm} w(m,k) \left\| \sum_{j=m+\lfloor kh_2 \rfloor + 1}^{m+k} H(X_j, \theta_0) \right\|_A$$

$$\leqslant \frac{1}{\sqrt{T}} \sup_{k \geqslant Tm} \frac{k}{m} \rho\left(\frac{k}{m}\right) \sup_{k \geqslant Tm} \frac{\sqrt{Tm}}{k} \left\| \sum_{j=m+\lfloor kh_2 \rfloor + 1}^{m+k} H(X_j, \theta_0) \right\|_A$$

$$= \frac{1}{\sqrt{T}} O_P(1) \xrightarrow{P} 0 \tag{10}$$

as $T \to \infty$ (uniformly in m). Analogously, we get for the Wiener process as in RC.1 for $T \to \infty$ uniformly in m by the law of iterated logarithm

$$\sup_{t \geqslant T} \rho(t) \left\| W_1(1+t) - W_1(1+th_2) \right\|_A = o_P(1).$$

By RC.1 and RC.2 it holds for $m \to \infty$ and T fixed

$$\sup_{k \geqslant \tau m} \left\| w(m, \min(k, mT)) \sum_{j=m+\lfloor \min((k,mT)h_2)\rfloor+1}^{m+\min(k,mT)} H(X_j, \theta_0) \right.$$

$$\left. -w(m,k)\frac{k-\lfloor kh_2 \rfloor}{m}B(\theta_0)\sum_{j=1}^{m}G(X_j,\theta_0) \right\|_A$$

$$\xrightarrow{D} \sup_{t \geqslant \tau} \|\rho(\min(t,T))(W_1(1+\min(t,T)) - W_1(1+\min(t,T)h_2))$$

$$-t(1-h_2)\rho(t)W_2(1)\|_A.$$

A careful combination of the above results yields assertion (b)(i).

The proof of (ii) is very close, so we concentrate on the difference. First, by an argument similar to (8) it is sufficient to consider the limit distribution of

$$\sup_{1 \leqslant k < N(m)} w(m,k) \max_{0 \leqslant i < k} \left\| \sum_{j=m+i+1}^{m+k} H(X_j, \theta_0) - \frac{k-i}{m}B(\theta_0)\sum_{j=1}^{m}G(X_j,\theta_0) \right\|_A.$$

The main difference in the proof is the following step:

$$\sup_{1 \leqslant k < N(m)} \min\left(\frac{1}{m^{\frac{1}{2}-\gamma}k^\gamma}, \frac{m^{1/2}}{k}\right) \max_{1 \leqslant i < k} \left\| \sum_{t=m+i+1}^{m+k} H(X_t, \widehat{\theta}_m) \right.$$

$$\left. - \left(\sum_{t=m+i+1}^{m+k} H(X_t, \theta_0) - \frac{k-i}{m}B(\theta_0)\sum_{j=1}^{m}G(X_j,\theta_0) \right) \right\|$$

$$\leqslant 2 \sup_{1 \leqslant i < N(m)} \min\left(\frac{1}{m^{\frac{1}{2}-\gamma}i^\gamma}, \frac{m^{1/2}}{i}\right) \left\| \sum_{t=m+1}^{m+i} H(X_t, \widehat{\theta}_m) \right.$$

$$\left. - \left(\sum_{t=m+1}^{m+i} H(X_t, \theta_0) - \frac{i}{m}B(\theta_0)\sum_{j=1}^{m}G(X_j,\theta_0) \right) \right\|$$

$$= o_P(1).$$

A similar argument can be used to obtain analogously to (9)

$$\sup_{1 \leqslant k < \tau m} w(m,k) \max_{0 \leqslant i < k} \left\| \sum_{j=m+i+1}^{m+k} H(X_j, \theta_0) - \frac{k-i}{m}B(\theta_0)\sum_{j=1}^{m}G(X_j,\theta_0) \right\|_A = o_P(1)$$

as $\tau \to 0$ uniformly in m.

A major difference in the proofs occurs when dealing with $k \geqslant Tm$ as in (10),

where the key difference is the following:

$$\sup_{k \geqslant Tm} \frac{\sqrt{m}}{k} \max_{1 \leqslant i \leqslant k} \left\| \sum_{j=m+1}^{m+i} H(X_j, \theta_0) \right\|$$

$$\leqslant \frac{1}{T} \max_{1 \leqslant i \leqslant m} \frac{1}{\sqrt{m}} \left\| \sum_{j=m+1}^{m+i} H(X_j, \theta_0) \right\| + \frac{1}{\sqrt{T}} \max_{m \leqslant i \leqslant \sqrt{T}m} \frac{\sqrt{m}}{\sqrt{T}m} \left\| \sum_{j=m+1}^{m+i} H(X_j, \theta_0) \right\|$$

$$+ \frac{1}{T^{1/4}} \max_{\sqrt{T}m \leqslant i \leqslant Tm} \frac{T^{1/4}\sqrt{m}}{Tm} \left\| \sum_{j=m+1}^{m+i} H(X_j, \theta_0) \right\|$$

$$+ \frac{1}{\sqrt{T}} \max_{i \geqslant Tm} \frac{\sqrt{Tm}}{i} \left\| \sum_{j=m+1}^{m+i} H(X_j, \theta_0) \right\| = T^{-1/4} O_P(1),$$

where this follows from RC.1 (b) for the first summand and RC.1 (c) for the last three summands.

The other parts of the proof are analogously to the arguments for (i). ☐

Proof of Theorem 2.2. By Hušková and Koubková (2005), proof of Theorem 1, it holds

$$\{W_1(1+t) - tW_2(1)\} \overset{\mathcal{D}}{=} \left\{ (1+t) W \left(\frac{t}{1+t} \right) \right\},$$

hence

$$\sup_{t>0} \left\| \frac{W_1(1+t) - tW_2(t) - (W_1(1+th_2) - th_2 W_2(1))}{(1+t) \left(\frac{t}{1+t} \right)^{\gamma}} \right\|_A$$

$$\overset{D}{=} \sup_{t>0} \left\| \left| \frac{(1+t)W_1 \left(\frac{t}{1+t} \right) - (1+th_2)W_1 \left(\frac{th_2}{1+th_2} \right)}{(1+t) \left(\frac{t}{1+t} \right)^{\gamma}} \right| \right\|_A$$

$$\overset{D}{=} \sup_{0<r<1} \left\| \left| \frac{W_1(r)}{r^{\gamma}} - (1 - (1-h_2)r) \frac{W_1 \left(\frac{h_2 r}{1-(1-h_2)r} \right)}{r^{\gamma}} \right| \right\|_A,$$

where the last equality follows from the substitution $r = t/(1+t)$, proving (a). The proof of (b) is analogous, where in the last step the mapping $(s,t) \mapsto (s/(1-s), t/(1-t))$ has been used. ☐

Proof of Theorem 2.3. The proof is analogous to the proof of Theorem 2.1 and therefore omitted. ☐

Proof of Theorem 2.4. As in the proof of Theorem 2.1 by RC.6 it is sufficient

to consider

$$
\sup_{1 \leqslant k < N(m)+1} w_M(h_4, k) \left\| \sum_{i=m+k-h_4+1}^{m+k} H(X_i, \theta_0) - \frac{h_4}{m} B(\theta_0) \sum_{j=1}^{m} G(X_j, \theta_0) \right\|_A
$$

$$
= \sup_{1 \leqslant k < N(m)+1} w_M(h_4, k) \left\| \sum_{i=m+k-h_4+1}^{m+k} H(X_i, \theta_0) \right\|_A + o_P(1)
$$

$$
\overset{\mathcal{D}}{=} \sup_{1 \leqslant k < N(m)+1} h_4^{-1/2} \rho(k/h_4) \left\| \sum_{i=k+1}^{k+h_4} H(X_i, \theta_0) \right\|_A + o_P(1)
$$

By RC.5 we get

$$
\sup_{1 \leqslant k < N(m)+1} h_4^{-1/2} \rho(k/h_4) \left\| \sum_{i=k+1}^{k+h_4} H(X_i, \theta_0) \right\|_A
$$

$$
= \sup_{1 \leqslant k < N(m)+1} h_4^{-1/2} \rho(k/h_4) \left\| W_1(k+h_4) - W_1(k) \right\|_A
$$

$$
+ O_P \left(\max_{1 \leqslant k < N(m)+1} \frac{(k+h_4)^{1/\nu}}{h_4^{1/2}} \min \left(1, \left(\frac{h_4}{k} \right)^{1/\nu} \right) \right)
$$

$$
= \sup_{1 \leqslant k < N(m)+1} h_4^{-1/2} \rho(k/h_4) \left\| W_1(k+h_4) - W_1(k) \right\|_A + o_P(1).
$$

The first summand converges to

$$
\sup_{0 < t < \infty} \rho(t) \| W_1(t+1) - W_1(t) \|_A,
$$

which is well-defined by Theorem 1.2.1 in Csörgö and Révész (1981). This completes the proof. □

Proof of Corollary 2.1. The proof is analogous to the proof of Theorem 2.4, where the supremum is taken over a finite stretch after substituting k with k/h_4. □

Proof of Theorem 2.5. For better readibility we use $h = h_4$ in this proof. Some calculations yield

$$
\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_0 = \left(\sum_{i=1}^{m} \boldsymbol{x}_i^T \boldsymbol{x}_i \right)^{-1} \sum_{i=1}^{m} \boldsymbol{x}_i X_i - \boldsymbol{\beta}_0 = \boldsymbol{C}_m^{-1} \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{x}_i \epsilon_i.
$$

With $\widehat{\epsilon}_i = X_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_m$ this implies

$$
\sum_{i=m+k-h+1}^{m+k} \widehat{\epsilon}_i = \sum_{i=m+k-h+1}^{m+k} \epsilon_i - \left(\sum_{i=m+k-h+1}^{m+k} \boldsymbol{x}_i \right)^T \boldsymbol{C}_m^{-1} \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{x}_i \epsilon_i, \quad (11)
$$

where $\boldsymbol{C}_m = \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^T$. Denote by $\mathbf{c}_1$ the first column of $\boldsymbol{C}$, so that by $\boldsymbol{c}_1^T\boldsymbol{C}^{-1}\sum_{i=1}^{m}\boldsymbol{x}_i\epsilon_i = \sum_{i=1}^{m}\epsilon_i$ it holds

$$\sum_{i=m+k-h+1}^{m+k}\widehat{\epsilon}_i - \left(\sum_{i=m+k-h+1}^{m+k}\epsilon_i - \frac{h}{m}\sum_{i=1}^{m}\epsilon_i\right)$$

$$= \left(\frac{1}{\sqrt{m}}\left(\left(\sum_{i=m+k-h+1}^{m+k}\boldsymbol{x}_i\right)^T - h\boldsymbol{c}_1^T\right)\left(\boldsymbol{C}_m^{-1} - \boldsymbol{C}^{-1}\right)\right. \tag{12}$$

$$+ \frac{1}{\sqrt{m}}\left(\left(\sum_{i=m+k-h+1}^{m+k}\boldsymbol{x}_i\right)^T - h\boldsymbol{c}_1^T\right)\boldsymbol{C}^{-1}$$

$$\left. + \frac{h}{\sqrt{m}}\boldsymbol{c}_1^T\left(\boldsymbol{C}_m^{-1} - \boldsymbol{C}^{-1}\right)\right)\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\boldsymbol{x}_i\epsilon_i. \tag{13}$$

Because $\{\epsilon_i, 1 \leqslant i < \infty\}$ and $\{\boldsymbol{x}_i, 1 \leqslant i < \infty\}$ are independent, it holds

$$\sum_{i=1}^{m}\boldsymbol{x}_i\epsilon_i = O_P(\sqrt{m}). \tag{14}$$

Furthermore, observe that by (6)

$$\sup_{1\leqslant k<N(m)}\min\left(\frac{1}{\sqrt{h}}, \frac{1}{h^{\frac{1}{2}-\frac{1}{\nu}}k^{\frac{1}{\nu}}}\right)\frac{h}{\sqrt{m}}\boldsymbol{c}_1^T\left(\boldsymbol{C}_m^{-1} - \boldsymbol{C}^{-1}\right)$$

$$\leqslant \sqrt{\frac{h}{m}}\boldsymbol{c}_1^T\left(\boldsymbol{C}_m^{-1} - \boldsymbol{C}^{-1}\right) = o_P(1). \tag{15}$$

By the stationarity of the regressors $\boldsymbol{x}_i$ we obtain

$$\sup_{1\leqslant k\leqslant Nh}\frac{1}{\sqrt{m}}\min\left(\frac{1}{\sqrt{h}}, \frac{1}{h^{\frac{1}{2}-\frac{1}{\nu}}k^{\frac{1}{\nu}}}\right)\left|\sum_{i=m+k-h+1}^{m+k}\boldsymbol{x}_i - h\boldsymbol{c}_1\right|$$

$$\stackrel{D}{=} \sup_{1\leqslant k\leqslant Nh}\frac{1}{\sqrt{m}}\min\left(\frac{1}{\sqrt{h}}, \frac{1}{h^{\frac{1}{2}-\frac{1}{\nu}}k^{\frac{1}{\nu}}}\right)\left|\sum_{i=k+1}^{k+h}\boldsymbol{x}_i - h\boldsymbol{c}_1\right|$$

$$\leqslant \sqrt{\frac{h}{m}}\sup_{1\leqslant k\leqslant Nh}\frac{1}{h}\left|\sum_{i=k+1}^{k+h}\boldsymbol{x}_i - h\boldsymbol{c}_1\right|$$

$$\leqslant \sup_{1\leqslant k\leqslant Nh}\frac{1}{h}\left|\sum_{i=1}^{k+h}\boldsymbol{x}_i - (k+h)\boldsymbol{c}_1\right| + \sup_{1\leqslant k\leqslant Nh}\frac{1}{h}\left|\sum_{i=1}^{k}\boldsymbol{x}_i - k\boldsymbol{c}_1\right|$$

$$\leqslant (N+1)\sup_{1\leqslant k\leqslant Nh}\frac{1}{Nh+h}\left|\sum_{i=1}^{k+h}\boldsymbol{x}_i - (k+h)\boldsymbol{c}_1\right| + N\sup_{1\leqslant k\leqslant Nh}\frac{1}{Nh}\left|\sum_{i=1}^{k}\boldsymbol{x}_i - k\boldsymbol{c}_1\right|$$

$$= o_P(1).$$

Putting this together with (13), (14) and (15) yields assertion (a).

For (b) it is sufficient to consider

$$
\sup_{k>Nh} \min\left(\frac{1}{\sqrt{h}}, \frac{1}{h^{\frac{1}{2}-\frac{1}{\nu}}k^{\frac{1}{\nu}}}\right) \left|\sum_{i=k+1}^{k+h} \boldsymbol{x}_i - h\boldsymbol{c}_1\right|
$$

$$
= \sup_{k>Nh} \min\left(\frac{1}{\sqrt{h}}, \frac{1}{h^{\frac{1}{2}-\frac{1}{\nu}}k^{\frac{1}{\nu}}}\right) \left(|\boldsymbol{W}_1(k+h) - \boldsymbol{W}_1(k)| + O_P\left((k+h)^{\frac{1}{\nu}}\right)\right)
$$

$$
\leqslant \sup_{k>Nh} \left(\frac{h}{k}\right)^{\frac{1}{\nu}} \left|\boldsymbol{W}_1\left(\frac{k}{h}+1\right) - \boldsymbol{W}_1\left(\frac{k}{h}\right)\right| + O_P(1) \sup_{k>Nh} \left(\frac{h}{k}\right)^{\frac{1}{\nu}} \frac{(k+h)^{\frac{1}{\nu}}}{\sqrt{h}}
$$

$$
\leqslant \sup_{t>N} t^{-\frac{1}{\nu}} \|\boldsymbol{W}_1(t+1) - \boldsymbol{W}_1(t)\| + o_P(1) = o_P(1)
$$

by Theorem 1.2.1 in Csörgö and Révész (1981), RC. 4a) and c). $\qquad\square$

Proof of Theorem 2.6. We start with the proof for the mMOSUM under RC.7 (d). Choose $\widetilde{k} = \lfloor x_0 m \rfloor$. If $\lfloor \widetilde{k}h_2 \rfloor < k^*$, then by RC.7 (a) and (b), it holds

$$
\sum_{t=m+\lfloor\widetilde{k}h_2\rfloor+1}^{m+\widetilde{k}} H(X_t, \widehat{\theta}_m) = \sum_{t=m+\lfloor\widetilde{k}h_2\rfloor+1}^{m+k^*} H(X_t, \widehat{\theta}_m) + \sum_{t=m+k^*+1}^{m+\widetilde{k}} H(X_t, \widehat{\theta}_m)
$$

$$
= m((x_0 - k^*/m) E_H + o_P(1)) \geqslant m\left((x_0 - \lambda)\, E_H + o_P(1)\right),
$$

from which we derive by RC.7 (c) and (d)

$$
w(m,\widetilde{k})\Gamma_2(m,\widetilde{k},h_2) \geqslant \sqrt{m}\,\rho\,(x_0 + o(1))\,(x_0 - \lambda)\left(\|E_H\|_A + o_P(1)\right) \xrightarrow{P} \infty.
$$

Similarly, if $\lfloor \widetilde{k}h_2 \rfloor \geqslant k^*$, then

$$
w(m,\widetilde{k})\Gamma_2(m,\widetilde{k},h_2) \geqslant \sqrt{m}\,\rho\,(x_0 + o(1))\,x_0(1 - h_2)\left(\|E_H\|_A + o_P(1)\right) \xrightarrow{P} \infty.
$$

In case of RC.7 (c) and (e) we get similarly by the choice $\widetilde{k} = \lceil k^*/h_2 \rceil$

$$
\sup_{k\geqslant 1} w(m,k) \left\|\sum_{m+\lfloor kh_2\rfloor+1}^{m+k} H(X_t, \widehat{\theta}_m)\right\|_A
$$

$$
\geqslant \frac{\widetilde{k}}{m}\,\rho\left(\frac{\widetilde{k}}{m}\right) m\,(1 - h_2)\left(\|E_H\|_A + o_P(1)\right) \xrightarrow{P} \infty.
$$

The proof of the Page-CUSUM ($j = 3$) follows from this by choosing $i = \lfloor kh_2 \rfloor$ or from the proof for the CUSUM (Kirch and Kamgaing (2015)) by choosing $i = 1$. For the MOSUM statistics we get analogously with the choice $\widetilde{k} = \lfloor x_0 h_4 \rfloor$ under RC.7(c) and (d)

$$
w(h_4,\widetilde{k})\Gamma_4(m,\widetilde{k},h_4) \geqslant \sqrt{h_4}\rho\,(x_0 + o(1))\min(x_0 - \lambda, 1)\left(\|E_H\|_A + o_P(1)\right) \xrightarrow{P} \infty,
$$

completing the proof. $\qquad\square$

References

E. Andreou and E. Ghysels. Monitoring disruptions in financial markets. *Journal of Econometrics*, 135:77–124, 2006. MR2328397

A. Aue and L. Horváth. Delay time in sequential detection of change. *Statistics and Probability Letters*, 67:221–231, 2004. MR2053524

A. Aue, L. Horváth, M. Husková, and P. Kokoszka. Change-point monitoring in linear models. *Econometrics Journal*, 9(3):373–403, 2006. MR2328569

A. Aue, L. Horváth, M. Kühn, and J. Steinebach. On the reaction time of moving sum detectors. *Journal of Statistical Planning and Inference*, 27(2):126–140, 2008. MR2911844

A. Aue, S. Hörmann, L. Horváth, and M. Hušková. Sequential testing for the stability of portfolio betas. *Econometric Theory*, 2011.

I. Berkes, E. Gombay, L. Horváth , and P. Kokoszka. Sequential change-point detection in GARCH(p,q) models. *Econometric theory*, 20:1140–1167, 2004. MR2101953

Z. Chen and Z. Tian. Modified procedures for change point monitoring in linear models. *Mathematics and computers in simulation*, 81(1):62–75, 2010. MR2729128

C.-S.J. Chu, M. Stinchcombe, and H. White. Monitoring structural change. *Econometrica*, 64:1045–1065, 1996.

G. Ciupera. Two tests for sequential detection of a change-point in a nonlinear model. *Journal of Statistical Planning and Inference*, 143(10):1719–1743, 2013. MR3082229

M. Csörgö and P. Révész. *Strong Approximations in Probability and Statistics*. Academic Press, 1981. MR0666546

B. Eichinger and C. Kirch. A mosum procedure for the estimation of multiple random change points. *Bernoulli*, 24(1):526–564, 2018. MR3706768

S. Fremdt. Page's sequential procedure for change-point detection in time series regression. *Statistics: A Journal of Theoretical and Applied Statistics*, 48(4):1–28, 2014. MR3304372

R. Fried and M. Imhoff. On the online detection of monotonic trends in time series. *Biometrical Journal*, 46:90–102, 2004. MR2055450

L. Horváth, M. Hušková, P. Kokoszka, and J. Steinebach. Monitoring changes in linear models. *Journal of statistical and inference*, 126(1):225–251, 2004. MR2090695

L. Horváth, M. Kühn, and J. Steinebach. On the performance of the fluctuation test for structural change. *Sequential Analysis*, 27(2):126–140, 2008. MR2422958

M. Hušková and C. Kirch. Bootstrapping sequential change-point tests for linear regression. *Metrika*, 75(5):673–708, 2012. MR2946654

M. Hušková and A. Koubková. Monitoring jump changes in linear models. *Journal on Statistical Research*, 39:51–70, 2005. MR2256552

C. Kirch. Bootstrapping sequential change-point tests. *Sequential Analysis*, 27(3):330–349, 2008. MR2446906

C. Kirch and J. Tadjuidje Kamgaing. Testing for parameter stability in nonlinear autoregressive models. *Journal of Time Series Analysis*, 33:365–385, 2012. MR2915090

C. Kirch and J. Tadjuidje Kamgaing. On the use of estimating functions in monitoring time series for change points. *Journal of Statistical Planning and Inference*, 161:25–49, 2015. MR3316549

C. Kirch and J. Tadjuidje Kamgaing. Detection of change points in discrete-valued time series. In Davis, R.A., Holan, S.A., Lund, R.B., and Ravishanker, N., editors, *Handbook of Discrete Valued Time series*, pages 219–244. CRC Press, 2016. MR3699407

C. Kirch, B. Muhsal, and H. Ombao. Detection of changes in multivariate time series with application to EEG data. *Journal of the American Statistical Association*, 110:1197–1216, 2015. MR3420695

S. Weber. *Change-Point Procedures for Multivariate Dependent Data*. PhD thesis, Karlsruhe Institute of Technology (KIT), 2017. URN: urn:nbn:de:swb:90-689812.

A. Zeileis. Implementing a class of structural change tests: An econometric computing approach. *Computational Statistics & Data Analysis*, 50(11):2987–3008, 2006. MR2239654

A. Zeileis, F. Leisch, K. Hornik, and C. Kleiber. strucchange: An R package for testing for structural change in linear regression models. *Journal of Statistical Software*, 7(2):1–38, 2002. URL http://www.jstatsoft.org/v07/i02/.

A. Zeileis, A. Shah, and I. Patnaik. Testing, monitoring, and dating structural changes in exchange rate regimes. *Computational Statistics & Data Analysis*, 54(6):1696–1706, 2010. MR2600868