Biometrika (2024), **111**, 2, *pp.* 367–391 https://doi.org/10.1093/biomet/asad079
Advance Access publication 21 December 2023

The state of cumulative sum sequential changepoint testing 70 years after Page

By ALEXANDER AUE

Department of Statistics, University of California, Davis, One Shields Avenue, Davis, California 95616, U.S.A.
aaue@ucdavis.edu

and CLAUDIA KIRCH

Department of Mathematics, Otto-von-Guericke University, Universitätsplatz 2, 39106 Magdeburg, Germany
claudia.kirch@ovgu.de

Summary

Quality control charts aim at raising an alarm as soon as sequentially obtained observations of an underlying random process no longer seem to be within stochastic fluctuations prescribed by an in-control scenario. Such random processes can often be modelled using the concept of stationarity, or even independence as in most classical works. An important out-of-control scenario is the changepoint alternative, for which the distribution of the process changes at an unknown point in time. In his seminal 1954 *Biometrika* paper, E. S. Page introduced the famous cumulative sum control charts for changepoint monitoring. Innovatively, decision rules based on cumulative sum procedures took the full history of the process into account, whereas previous procedures were based only on a fixed and typically small number of the most recent observations. The extreme case of using only the most recent observation, often referred to as the Shewhart chart, is more akin to serial outlier than changepoint detection. Page's cumulative sum approach, introduced seven decades ago, is ubiquitous in modern changepoint analysis, and his original paper has led to a multitude of follow-up papers in different research communities. This review is focused on a particular subfield of this research, namely nonparametric sequential, or online, changepoint tests that are constructed to maintain a desired Type-1 error as opposed to the more traditional approach seeking to minimize the average run length of the procedures. Such tests have originated at the intersection of econometrics and statistics. We trace the development of these tests and highlight their properties, mostly using a simple location model for clarity of exposition, but we also review more complex situations such as regression and time series models.

Some key words: Changepoint analysis; Cumulative sum; Monitoring procedure; Sequential testing.

1. Introduction

Control charts have been a staple of industrial quality control research ever since the pioneering work of Shewhart in the 1920s and 1930s. This work, summarized in the monograph by Shewhart (1931), includes first simple versions of sequential changepoint monitoring

procedures endowed with empirical decision 'three-σ' rules. Significant theoretical advancements were made by Wald (1947), in particular through the provision of the sequential likelihood ratio testing framework. This research set the scene for Page's seminal work (Page, 1954). Published 70 years ago, the paper remains a central piece for modern statistical applications, especially those connected to the expansive field of changepoint analysis. The most important innovation of Page (1954) was the introduction of a cumulative sum, CUSUM, statistic that kept track of the entire history of an observed process and not only of a small window containing the most recent data points, as hitherto customary for control charts. Variants of Page's CUSUM are now found in countless papers, addressing modern changepoint problems for ever more complex data, both with sequential (online) and a posteriori (off-line) procedures. Examples include machine learning for data streams (Krawczyk et al., 2017), where gradual changepoints are referred to as concept drift, remote sensing of spatial data of various resolutions to generate rapid information on forest disturbances such as fire and insect attacks (Olsson et al., 2016), monitoring of ambient air pollution (Dienes & Aue, 2014), as well as more generally the growing fields of precision agriculture (Milella et al., 2019; Kumar & Purbey, 2021) and precision medicine (Benincasa et al., 2019), to name a few.

The original works of Shewhart and Page were concerned with the sequential setting where observations are obtained one by one, leading to a monitoring process. After each new observation, a decision rule is applied to decide whether a changepoint has likely occurred. If so, an alarm is triggered and monitoring is stopped. More concisely, if the observations up to the current time-point k are recorded as $X_1, \ldots, X_k$, and if $S'_0 = 0$ and

$$S'_k = \max\{S'_{k-1} + X_k, 0\}, \qquad k \geqslant 1, \tag{1}$$

then one raises an alarm after the kth observation if $S'_k \geqslant h$ for an appropriate $h = h_k$. This implements a test against a one-sided alternative if there is an increase in the mean of the observations. The quantity S'_k is the celebrated Page CUSUM. Page (1954) himself made the connection to Wald's sequential probability ratio test (Wald, 1947), noting that this decision rule determines a sequence of Wald sequential tests with boundaries at 0 and h, and initial score 0. The test is repeated when the previous test ends on the lower boundary and an alarm is raised when the upper boundary is reached or exceeded. Many of the theoretical properties of Page's prescient work were only derived in the decades following the paper's publication. A series of papers, starting with Lorden (1971), derived the optimality of Page's CUSUM for independent observations in the sense that it produces the smallest average run length. The work of Lorden (1971) was continued by Moustakides (1986) and Ritov (1990), and extends to parametric models. The interested reader may find further generalizations of this approach in the monographs of Tartakovsky et al. (2014) and Tartakovsky (2019) and the surveys by Shiryaev (2010), Polunchenko & Tartakovsky (2012) and Xie et al. (2021), as the development of control charts based on optimal stopping continues to be an active area of research.

Now, typically, interest is in procedures minimizing the detection delay if a change occurs under less conservative constraints in the no-change case, such as requiring lower bounds on the time to false alarm; see Shiryaev (1961) and Roberts (1966, 2000).

In the past three decades or so, the focus of research has shifted from online to off-line procedures. In the off-line situation, the full dataset is available for analysis at the outset of the investigation, and can be utilized to test for one or more changepoints and estimate their location(s). Some of the latest incarnations of this methodology cast the multiple changepoint problem as a model selection procedure. Since a review of this literature is

beyond the scope of this paper, we refer to the recent survey articles by Aue & Horváth (2013), Truong et al. (2020) and Cho & Kirch (2021) for a wealth of references. Suffice it to say here that many of the proposed methods in these settings are based on some version of the CUSUM statistics in (1), often of the type

$$S_k - \frac{k}{n} S_n, \tag{2}$$

where $S_k = X_1 + \cdots + X_k$ $(k = 1, \ldots, n)$. The CUSUM in (2) does not have the same resetting mechanism as the CUSUM in (1), where a maximum is taken with respect to zero. The CUSUM in (2) can also be derived from likelihood ratio theory under Gaussianity, see, for example, (1.4.27) and §2.1 of Csörgő & Horváth (1997), and so inherits some of the optimality properties in the off-line setting.

In the past few years, owing to new capabilities in data acquisition and the need to make well-informed decisions while data keeps being generated, sequential procedures have returned to the fore of pertinent research; see the references in the first paragraph of this section. The online approaches mentioned above, including the original proposal by Page (1954), however, rely on a strong parametric assumption that in particular comprises the knowledge of the in-control distribution and often even the out-of-control distribution. This assumption, which may have been reasonable for the simple industrial production data Shewhart and Page had in mind, is difficult to maintain in modern applications. It is therefore more relevant to assume that in practice the in-control and out-of-control distributions are unknown and are to be estimated from previous in-control data. Even in the simplest possible case, where it is known that the change occurs in the expectation, the extra estimation step leads to additional statistical uncertainty that needs to be accounted for. This motivated Chu et al. (1996) to develop asymptotic theory for CUSUM monitoring schemes that integrate such an estimation step, an approach that has since seen further traction in literature at the intersection of statistics and econometrics, much of which will be discussed in the discourse of this review paper. The major novelty of the approach in Chu et al. (1996) was the provision of a history of data that can be used to calibrate to a no-change hypothesis without the need for parametric assumptions or known in-control distributions. The size of this history, say m, can be interpreted as a quantity representing confidence in understanding the in-control distribution. The off-line CUSUM in (2) can be adjusted to the online version

$$\Psi(m, k) = (S_{m+k} - S_m) - \frac{k}{m} S_m, \qquad k \geqslant 1, \tag{3}$$

where now asymptotic considerations may be carried out letting $m \to \infty$. The resulting procedures are nonparametric and control the false detection rate. The remainder of this survey discusses the main thrust of the research related to Chu et al. (1996) and the CUSUM in (3), including detailed background and intuition in §2. Extensions of these results are briefly reviewed in §3.

2. Online statistical testing for changepoints

2.1. *Background*

Over the past 20 years a modified approach to sequential changepoint methodology has gained ground. In contrast to other sequential methods this approach, asymptotically,

controls the false alarm rate at a given level rather than controlling the average run length of the procedure. The modified approach also does not assume knowledge of the in-control parameters of the underlying model. Because approaches based on controlling the average run length in the no-change situation do not control the false alarm rate, they are not statistical tests in the usual sense. On the other hand, controlling the false alarm rate comes at the cost of a longer detection delay, the sequential analogue of a Type-II error. Consequently, if a quick detection is more important than avoiding false alarms, approaches controlling the average run length are better suited. The opposite is true in situations where a quick detection is less important than statistical guarantees that the alarm is indeed a true positive in the usual testing sense.

This new way of conducting sequential tests originated with Chu et al. (1996), who made explicit mention of the work of Page (1954). Chu et al. (1996) were primarily interested in a linear regression framework to assess in real time the structural stability, in the sense of having no changepoints, of economic systems. As a derivative of Page (1954), this paper has spawned extensive research on its own. In a certain sense the tests proposed in this paper are a simplification of Wald and Page, as in today's applications it is often reasonable to assume that the data acquisition process is free of charge and therefore the monitoring of a process in the no-change scenario does not have to terminate due to accumulated expenses.

This section will lay out the general idea behind these methods for the simple location model where a change occurs in the expectation of otherwise independent and identically distributed data with unknown distribution satisfying certain moment assumptions. Extensions to more complex settings such as regression and time series models are postponed to § 3.

2.2. *The use of training data for asymptotic testing*

Classic asymptotic testing theory relies on the sample size to grow to infinity in order to identify inherent statistical regularities. This paradigm is not available in a sequential set-up where monitoring of the data stream continues until an alarm is raised. An additional problem arises from the fact that the in-control parameters are not known a priori, but can only be estimated from some initial in-control dataset. Both problems are solved by making use of a training or historical dataset with no change to estimate the in-control parameters. It is important to take the uncertainty stemming from this estimation into account in the procedure; see Remark 2 below. The existence of a training dataset is not an unreasonable assumption, as normally one will have observed a stable process for some time before starting to monitor for changes in the parameters. To elaborate, consider

$$X_i = \mu_i + \varepsilon_i, \qquad i \geqslant 1,$$

where $\{\varepsilon_i : i \geqslant 1\}$ is a sequence of centred independent and identically distributed random variables with finite second moments. In the in-control situation, i.e., the sequential null hypothesis,

$$\mu_i = \mu_0, \qquad i \geqslant 1, \tag{4}$$

for some unknown in-control parameter μ_0. In order to estimate this unknown in-control parameter, assume that there is training data of length m without contamination following

the in-control model

$$X_i = \mu_0 + \varepsilon_i \qquad (i = 1, \ldots, m).$$

One can think of this as the data that have already been observed in a controlled environment. In finance, this could be the data most recently observed and used to describe the current state of the market. Such an estimate is only useful in subsequent applications as long as the state of the market does not change too much and the parameters estimated from the training dataset still match the newly arriving data well enough. Therefore, it is of importance to have a procedure that raises an alarm as soon as this is no longer the case. In addition, the size m of the training dataset allows for rigorous asymptotic theory by letting m grow to infinity, while still allowing the monitoring to continue possibly forever.

As an out-of-control model, consider the sequential alternative for which the expected value changes according to

$$\mu_i = \mu_0, \quad i \leqslant m + k^*, \quad \text{but} \quad \mu_i = \mu_0 + d_m \neq \mu_0, \quad i > m + k^*,$$

where $m + k^*$ is an unknown point in time. The expectation after the change may depend on the size m of the training data to include local changes for which $d_m \to 0$ as $m \to \infty$, sufficiently slowly. Putting the above together, we consider the location model

$$X_i = \mu_0 + d_m 1_{\{i > m+k^*\}} + \varepsilon_i, \tag{5}$$

where $\{\varepsilon_i : i \geqslant 1\}$ is an independent and identically distributed sequence of centred random variables with existing nonzero variance. The observations $\{X_i : 1 \leqslant i \leqslant m\}$ are called the *training dataset*, while the *monitoring period* consists of the variables $\{X_i : i > m\}$.

The simplest monitoring statistics compare the sample mean of the available observations of the monitoring period with the sample mean of the observations from the training dataset, leading to

$$\Psi(m, k) = \sum_{i=m+1}^{m+k} (X_i - \bar{X}_m), \qquad \bar{X}_m = \frac{1}{m} \sum_{i=1}^{m} X_i;$$

see (3). Clearly, large absolute values indicate a change in expectation, while values fluctuating around zero do not. Therefore, it is natural to raise an alarm as soon as $|\Psi(m, k)|$ crosses a threshold. In order to control the familywise error rate, with respect to the tests at each location k, and since the fluctuations of $\Psi(m, k)$ increase with k, such a threshold will naturally depend on k, thus constituting a strictly positive critical curve, as opposed to a critical value as for classical a posteriori testing. Alternatively, one can consider a weight function $w(m, k) \geqslant 0$ and raise an alarm as soon as $w(m, k)|\Psi(m, k)| \geqslant m^{1/2} c_\alpha$, where c_α now plays the role of the classical critical value in a posteriori testing. We follow this latter approach as it allows for $w(m, k) = 0$ for certain values of k, which means that the time-point $m + k$ is excluded from the monitoring. In particular, it allows for a simultaneous treatment of both open-end procedures, where monitoring continues until infinity if no alarm is raised, and closed-end procedures, where monitoring stops after a prespecified time horizon. Because the derivation of the null asymptotics is based on a functional central limit theorem for the partial sum process $\{\sum_{j=1}^{k} X_j / \sqrt{m} : 1 \leqslant j \leqslant (N+1)m\}$ for some arbitrary but fixed constant N, the scaling with $\sqrt{m}$ is rather natural.

The following class of weight functions is popular in the literature because it leads to a nice limit distribution and it fulfils all necessary assumptions for the theoretical results to follow:

$$w_\gamma(m,k) = \left(1+\frac{k}{m}\right)^{\gamma-1}\left(\frac{k}{m}\right)^{-\gamma}, \qquad 0 \leqslant \gamma < \frac{1}{2}. \tag{6}$$

The parameter γ indexing the weight function $w_\gamma(m,k)$, and more generally its shape, determines the reaction time of the sequential test to deviations from the in-control scenario. Indeed, if one considers the decision at time-point k as a separate test then the familywise error rate is controlled at a global level α, where the shape of the curve determines how this α is spent among the locations and, consequently, also the associated power and detection delay for each of these locations. Previous literature on parametric sequential testing (Lan & DeMets, 1983; Gandy, 2009) calculates the corresponding critical curve from an increasing 'spending sequence' α_k, with the interpretation that the probability of a false rejection before time-point k occurs with probability at most α_k. In contrast, here, the shape of the curve is chosen via the weight function, thus directly influencing the locations with higher power, corresponding to a lower curve. Nevertheless, the familywise error rate is asymptotically controlled by scaling the curve with the appropriate asymptotic quantiles. Typically, the critical curves are increasing in k, where often two different curves cross each other exactly once. In that situation the lower curve leads to quicker rejections, such that the practitioner can choose a curve most suitable to the expected times of changes while still having asymptotic power one for other locations of the changepoint.

Summing up, for k such that $w(m,k) > 0$, an alarm is raised as soon as the monitoring statistic $\Psi(m,k)$ exceeds a critical curve given by $m^{1/2}c_\alpha/w(m,k)$. Otherwise, monitoring continues, possibly until infinity. Thus, the stopping time τ_m of the procedure is given by

$$\tau_m = \inf\{k \geqslant 1 : w(m,k)|\Psi(m,k)| > c_\alpha \sqrt{m}\},$$

where $\inf(\emptyset) = \infty$.

2.3. *Asymptotic size and power*

The set-up of the previous section allows the construction of sequential testing procedures for the null hypothesis of no change $H_0 : k^* = \infty$ against the alternative $H_1 : k^* < \infty$. These procedures asymptotically control the size in the sense that

$$\lim_{m\to\infty} P_{H_0}(\tau_m < \infty) = \lim_{m\to\infty} P_{H_0}\Big(\sup_{k\geqslant 1} w(m,k)|\Psi(m,k)| > c_\alpha\sqrt{m}\Big) = \alpha \tag{7}$$

for a suitable critical value c_α, where P_{H_0} denotes the probability under the null hypothesis specified by (4). Based on the asymptotic set-up including the training dataset, it is possible to derive the limit distribution of $\sup_{k\geqslant 1} w(m,k)|\Psi(m,k)|/\sqrt{m}$ under the null hypothesis. Thus, choosing the critical value c_α as the $(1-\alpha)$-quantile of this limit distribution achieves (7) in the same way as in classical a posteriori testing.

To this end, we impose the following assumptions on the weight functions, which are fulfilled by the class given in (6).

Assumption 1. Let $w(m,k) = \rho(k/m)$, where $\rho : (0,\infty) \to [0,\infty)$ is nonnegative and continuous if restricted to $(0, e_\rho] \cap \mathbb{R}$, where $e_\rho = \sup\{t > 0 : \rho(t) > 0\}$, with

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

Essentially, the weight function w relates to ρ via a rescaling with respect to m, a technique commonly used in asymptotic statistics. The other two assumptions are needed to control the asymptotic behaviour of the monitoring statistics at the beginning of the monitoring period and infinity, respectively. The above formulation includes both open-end procedures, for which $\sup\{t\colon \rho(t) > 0\} = \infty$, so that monitoring potentially continues forever, as well as closed-end procedures, for which $\rho(t) = 0$ for all $t > N$, so that monitoring stops after Nm observations.

Under these assumptions, we can prove the following theorem.

THEOREM 1. *Under the null hypothesis, i.e., for $X_i = \mu_0 + \varepsilon_i$, where $\{\varepsilon_i\}$ is an independent and identically distributed sequence with $\mathsf{E}\varepsilon_i = 0$ and $0 < \mathsf{Var}(\varepsilon_i) = \sigma^2 < \infty$, it holds under Assumption 1 that*

$$\frac{1}{(m\,\hat{\sigma}_m^2)^{1/2}} \sup_{k\geqslant 1} w(m,k)|\Psi(m,k)| \xrightarrow{\mathrm{D}} \sup_{0<t<1} \rho\left(\frac{t}{1-t}\right)\frac{|W(t)|}{1-t},$$

where $\{W(\cdot)\}$ is a standard Brownian motion and $\hat{\sigma}_m^2$ some consistent variance estimator $\hat{\sigma}_m^2 \xrightarrow{\mathrm{P}} \sigma^2$ calculated only from the training data.

Most notably, the limit distribution is a functional of a Brownian motion, where the supremum is taken only over the unit interval rather than the full positive line $\mathbb{R}_+$. This is possible due to the rescaling of time with respect to the size of the training dataset m that relates the weight function $w(\cdot,\cdot)$ to $\rho(\cdot)$. This is particularly useful for simulating the quantiles from the limit distribution. Using the $(1-\alpha)$-quantiles of the limit process, multiplied by $\hat{\sigma}_m$, as a critical value c_α guarantees that the sequential test asymptotically controls the probability of a false alarm in the sense of (7). A standard choice for $\hat{\sigma}_m^2$ is the empirical variance based on the training dataset

$$\hat{\sigma}_m^2 = \frac{1}{m-1}\sum_{j=1}^{m}(X_j - \bar{X}_m)^2. \tag{8}$$

Remark 1. It has been observed in the literature that the above monitoring scheme, in particular in combination with a γ close to $1/2$, tends to have too many false positives at the very first observations from the monitoring period, because the asymptotic distribution underestimates the fluctuations of $\sum_{j=m+1}^{m+k}(X_j - \bar{X}_m)$ for small values of k. This problem is easily solved by waiting for a few, say $\log(m)$, observations before actually starting with the monitoring. Indeed, the proof given below remains correct if one allows for a weight function with $w(m,k) = 0$ for $k \leqslant l_m$ with $l_m/m \to 0$, which is exactly the situation where monitoring only starts after l_m observations. This does indeed stabilize the small-sample size by avoiding false alarms at the very onset of the monitoring period; see Kirch & Weber (2018).

Sketch proof of Theorem 1. The proof is a special case of the proof of Theorem 2.1 of Kirch & Weber (2018) or Theorem 1 of Kirch & Stoehr (2022b), which expanded on the proof techniques first developed by Horváth et al. (2004).

The assertion with $\hat{\sigma}^2$ follows immediately from Slutsky's lemma and the assertion with σ^2 known. Thus, without loss of generality, let $\sigma^2 = 1$ be known.

To train intuition, we start with the proof for $\gamma = 0$ and the closed-end situation with $\rho(t) = 0$ for all $t > N$. It follows from the functional central limit theorem and continuity of ρ that

$$\begin{aligned}
&\frac{1}{\sqrt{m}} \sup_{k \geqslant 1} w(m,k)|\Psi(m,k)| \\
&\quad = \sup_{l_m < k \leqslant mN} \rho\left(\frac{k}{m}\right)\left|\frac{1}{\sqrt{m}} \sum_{i=m+1}^{m+k} \varepsilon_i - \frac{k}{m}\frac{1}{\sqrt{m}} \sum_{j=1}^{m} \varepsilon_j\right| \\
&\quad = \sup_{l_m/m < s \leqslant N} \rho\left(\frac{\lfloor ms \rfloor}{m}\right)\left|\frac{1}{\sqrt{m}} \sum_{i=m+1}^{m+\lfloor ms \rfloor} \varepsilon_i - \frac{\lfloor ms \rfloor}{m}\frac{1}{\sqrt{m}} \sum_{j=1}^{m} \varepsilon_j\right| \\
&\quad = \sup_{0 < s \leqslant N} \rho(s)\left|\frac{1}{\sqrt{m}} \sum_{i=m+1}^{m+\lfloor ms \rfloor} \varepsilon_i - s\frac{1}{\sqrt{m}} \sum_{j=1}^{m} \varepsilon_j\right| + o_P(1) \\
&\quad \xrightarrow{\mathrm{D}} \sup_{0 < s \leqslant N} \rho(s)|W(1+s) - (1+s)W(1)|,
\end{aligned}$$

because in this closed-end situation with $\rho(t) = 0$ for $t > N$, the latter expression is equal to $\sup_{s>0} \rho(s)|W(1+s) - (1+s)W(1)|$. A comparison of the covariance structure shows that

$$\{W(1+s) - (1+s)W(1) : s \geqslant 0\} \stackrel{\mathrm{D}}{=} \left\{(1+s)W\left(\frac{s}{1+s}\right) : s \geqslant 0\right\},$$

such that, by a substitution of $t = s/(1+s)$,

$$\begin{aligned}
\sup_{s>0} \rho(s)|W(1+s) - (1+s)W(1)| &\stackrel{\mathrm{D}}{=} \sup_{s>0} \rho(s)(1+s)\left|W\left(\frac{s}{1+s}\right)\right| \\
&= \sup_{0<t<1} \rho\left(\frac{t}{1-t}\right)\frac{|W(t)|}{1-t}. \qquad (9)
\end{aligned}$$

If Assumption 1(i) does not hold for $\gamma = 0$, but only $0 < \gamma < 1/2$, i.e., when $\lim_{t \to 0} \rho(t) = \infty$, the following cutting techniques at 0 are required to obtain the result. Indeed, for any $0 < \xi < 1$, analogously to the above we obtain

$$\frac{1}{\sqrt{m}} \sup_{\xi m \leqslant k \leqslant mN} w(m,k)|\Psi(m,k)| \xrightarrow{\mathrm{D}} \sup_{\xi \leqslant s \leqslant N} \rho(s)(1+s)\left|W\left(\frac{s}{1+s}\right)\right|. \qquad (10)$$

Furthermore, by the Hájek–Rényi inequality (see, e.g., Lin & Bai, 2010, 6.6.b), it holds for $\tilde{\gamma} > \gamma$ that

$$\begin{aligned}
&\frac{1}{\sqrt{m}} \sup_{1 \leqslant k < \xi m} w(m,k)|\Psi(m,k)| \\
&\quad \leqslant \max_{0<s<\xi} s^{\tilde{\gamma}}\rho(s)\left(\sup_{1 \leqslant k < m} \frac{1}{m^{1/2-\tilde{\gamma}}k^{\tilde{\gamma}}} \left| \sum_{i=m+1}^{m+k} \varepsilon_i \right| + \left| \frac{1}{\sqrt{m}} \sum_{j=1}^{m} \varepsilon_j \right| \right) \\
&\quad = O_P(1) \max_{0<s<\xi} s^{\tilde{\gamma}}\rho(s),
\end{aligned}$$

where the $O_P(1)$ is uniformly in m. By Assumption 1(i), it holds that $\max_{0<s<\xi} s^{\tilde{\gamma}}\rho(s) \to 0$ for $\xi \to 0$. Similar arguments show that $\sup_{s<\xi} \rho(s)(1+s)|W\{s/(1+s)\}| = o_P(1)$ for $\xi \to 0$. Carefully combining this with (10) shows that

$$\frac{1}{\sqrt{m}} \sup_{1 \leqslant k \leqslant mN} w(m,k)|\Psi(m,k)| \xrightarrow{\mathrm{D}} \sup_{1<s\leqslant N} \rho(s)(1+s)\left| W\left(\frac{s}{1+s}\right)\right|;$$

see also Lemma B.2 in the 2019 PhD thesis by C. Stoehr from the Otto-von-Guericke University, Magdeburg. To include the open-end procedure, we need to use similar cutting techniques at infinity, where the situation is slightly more complicated because the second term $(k/m)\sum_{j=1}^{m}\epsilon_j/\sqrt{m}$ is not asymptotically negligible. Fortunately, the dependence on k is only in the deterministic term k/m, while the random term $\sum_{j=1}^{m}\epsilon_j/\sqrt{m}$ does not depend on k such that the result can be obtained by a suitable functional central limit theorem in combination with the continuous mapping theorem. Indeed, the continuity follows from Assumption 1(ii). In the case of an open-end procedure, we get, for any $N > 0$, in combination with the above cutting techniques at 0 for $0 < \gamma < 1/2$,

$$\begin{aligned}
&\sup_{k\geqslant 1} w(m,k)\left| \frac{1}{\sqrt{m}} \sum_{i=m+1}^{m+\min(k,mN)} \varepsilon_i - \frac{k}{m}\frac{1}{\sqrt{m}}\sum_{j=1}^{m}\varepsilon_j \right| \\
&\qquad \xrightarrow{\mathrm{D}} \sup_{s>0} \rho(s)|W\{1+\min(s,N)\} - (1+s)W(1)|. \qquad (11)
\end{aligned}$$

Now, the Hájek–Rényi inequality, extended to an unbounded maximum, which is easily achieved by continuity of measures, yields

$$\begin{aligned}
&\left| \frac{1}{\sqrt{m}} \sup_{k\geqslant 1} w(m,k)|\Psi(m,k)| - \sup_{k\geqslant 1} w(m,k)\left| \frac{1}{\sqrt{m}} \sum_{i=m+1}^{m+\min(k,mN)} \varepsilon_i - \frac{k}{m}\frac{1}{\sqrt{m}}\sum_{j=1}^{m}\varepsilon_j \right| \right| \\
&\quad \leqslant \sup_{s\geqslant N} s\rho(s) \max_{k\geqslant mN} \frac{m^{1/2}}{k} \left| \sum_{i=mN+1}^{m+k} \varepsilon_i \right| \\
&\quad = O_P\left(\frac{1}{\sqrt{N}}\right) \sup_{s\geqslant N} s\rho(s),
\end{aligned}$$

where the O_P term is uniform in m. The right-hand side converges to 0 for $N \to \infty$ and an analogous assertion can also be shown for the difference between the limit in (11) and

$\sup_{s>0} \rho(s)|W(1+s)-(1+s)W(1)|$. A careful combination with (11) and (9) yields the assertion; see also Lemma B.2 in the 2019 PhD thesis by C. Stoehr from the Otto-von-Guericke University, Magdeburg. □

Remark 2. The proof shows that

$$\sup_{0<t<1} \rho\left(\frac{t}{1-t}\right)\frac{|W(t)|}{1-t} \stackrel{\mathrm{D}}{=} \sup_{s>0} \rho(s)|W(1+s)-(1+s)W(1)|.$$

If instead of estimating μ_0 by $\bar{X}_m$ the true value were used, the same arguments show that the limit is given by $\sup_{s>0} \rho(s)|W(1+s)-W(1)|$. Thus, the additional uncertainty stemming from the estimation of the unknown parameter is reflected in the additional term $sW(1)$ and taken into account by this procedure.

Furthermore, the above stopping procedure has asymptotic power one under weak assumptions on the alternative in combination with the weight functions.

THEOREM 2. *Let the alternative* $X_i = \mu_0 + d_m 1_{\{i>m+k^*\}} + \varepsilon_i$ *hold with* $m^{1/2}|d_m| \to \infty$, *where* $\{\varepsilon_i\}$ *is an independent ad identically distributed sequence with* $\mathsf{E}\varepsilon_i = 0$ *and* $0 < \mathsf{Var}(\varepsilon_i) = \sigma^2 < \infty$. *Additionally, for superlinear changes fulfilling* $k^*/m \to \infty$, *let* $\rho(\cdot)$ *as in Assumption* 1 *fulfil* $\liminf_{t\to\infty} t\rho(t) > 0$. *For sublinear and linear changes with* $\limsup_{m\to\infty} k^*/m \leqslant t_1$ *for some* t_1, *let* $\inf_{t\in[t_2,t_3]} \rho(t) \geqslant c > 0$ *for some* $t_1 < t_2 < t_3$. *Then,*

$$\frac{1}{\sqrt{m}} \sup_{k\geqslant 1} w(m,k)|\Psi(m,k)| \stackrel{\mathrm{P}}{\to} \infty,$$

such that the corresponding sequential test has asymptotic power one.

The quantity $m^{1/2}|d_m|$ can be thought of as the signal strength, where we require $m^{1/2}|d_m| \to \infty$ to obtain asymptotic power one. This shows that the magnitude of the change must be larger for good power behaviour if the length of the training dataset is smaller, similar to what is known in a posteriori testing. Indeed, a smaller training sample results in more uncertainty in the estimate $\bar{X}_m$ of the in-control mean μ_0, explaining the effect.

The above theorem contains both the situation of fixed alternatives with $d_m = d$ not depending on m as well as local alternatives with $d_m \to 0$, but not too slow. The consideration of local changepoints is a well-established asymptotic tool to get a better insight into the power of a test for small effects, relative to the sample size. This is related to asymptotic efficiency concepts for tests, which are known to give a better insight into the small-sample behaviour than fixed changes; see Aston & Kirch (2018) for a modern version in the context of high-dimensional testing.

Sketch proof of Theorem 2. For any $\widetilde{k} > \max(k^*, m)$, it holds that

$$\begin{aligned}\frac{1}{\sqrt{m}} w(m,\tilde{k})\Psi(m,\widetilde{k}) &= \rho\left(\frac{\widetilde{k}}{m}\right)\frac{\widetilde{k}-k^*}{m}m^{1/2}d_m + \frac{\widetilde{k}}{m}\rho\left(\frac{\widetilde{k}}{m}\right)\left(\frac{\sqrt{m}}{\widetilde{k}}\sum_{i=m+1}^{m+\widetilde{k}}\varepsilon_i - \frac{1}{\sqrt{m}}\sum_{j=1}^{m}\varepsilon_j\right)\\ &= \rho\left(\frac{\widetilde{k}}{m}\right)\frac{\widetilde{k}-k^*}{m}m^{1/2}d_m + O_P(1),\end{aligned}$$

as can be seen by calculating the variance of the sums. It remains to be shown that the absolute value of the first summand converges to infinity for a suitably chosen $\tilde{k}$. Indeed, for the superlinear change situation, a choice of $\tilde{k} = 2k^*$ achieves this as, by assumption, $\liminf_{t\to\infty}(k^*/m)\rho(2k^*/m) > 0$ and $m^{1/2}|d_m| \to \infty$. Similarly, for early changes, a choice of $\tilde{k} = \lfloor t_2 m \rfloor + 1$ achieves this as, for large enough m, by assumption, $\rho\{(\lfloor t_2 m \rfloor + 1)/m\}(\lfloor t_2 m \rfloor + 1 - k^*)/m \geqslant c\,(t_2 - t_1) > 0$ and $m^{1/2}|d_m| \to \infty$. □

For the weight function in (6) with $\gamma = 1/2$, standard distributional convergence no longer holds. Instead, the following extreme-value result (see Theorem 1.1 of Horváth et al., 2007) can be established:

$$a(\log m)\frac{1}{(m\hat{\sigma}_m^2)^{1/2}} \sup_{1\leqslant k\leqslant N(m)} w_{1/2}(m,k)|\Psi(m,k)| - b(\log(m)) \xrightarrow{\mathrm{D}} \Gamma. \tag{12}$$

Here

$$a(x) = (2\log x)^{1/2}, \qquad b(x) = 2\log x + \tfrac{1}{2}\log\log x - \tfrac{1}{2}\log\pi,$$

Γ is a Gumbel distribution with $P(\Gamma \leqslant t) = \exp\{-\exp(-t)\}$ and $N(m) = O(m^\lambda)$ with $1 \leqslant \lambda < \infty$ and $\liminf_{m\to\infty} N(m)/m > 0$. A similar procedure was considered by Aue & Kühn (2008). Such extreme-value asymptotics are known to often be slow.

2.4. *Asymptotic delay times*

The CUSUM monitoring procedure discussed in the previous section was shown to have asymptotic power one. A refinement of that statement concerns the derivation of the behaviour of its delay time, the time it takes for the statistic to raise an alarm after a changepoint has occurred. This gives important complementary information and allows in particular for a meaningful asymptotic comparison of different monitoring schemes with asymptotic power one.

In the remainder of this section, we consider model (5) with $\{\varepsilon_i : i \geqslant 1\}$ independent and identically distributed, $\mathsf{E}(\varepsilon_i) = 0$ and $0 < \mathsf{Var}(\varepsilon_i) < \infty$. In order to get meaningful insights into the stopping time distribution for finite samples by asymptotic theory, we need to make the changepoint dependent on the length of the training dataset m, the quantity converging to infinity for asymptotic considerations. Therefore, in the proofs, we consider changepoints of the form $k^* = \lfloor \lambda m^\beta \rfloor$ with $\lambda > 0$ and $\beta \geqslant 0$, both unknown. Here, effectively, β determines how early in the monitoring period, relative to the length of the training dataset, the changepoint occurs, which does influence the asymptotic behaviour. We distinguish three main cases, namely sublinear changes ($\beta < 1$), which model very early changes, linear ($\beta = 1$) changes, which model changes well within the monitoring period as multiples of m and superlinear ($\beta > 1$) changes, which model the behaviour for changes that occur late, i.e., after several, asymptotically increasing, multiples of m. In order to go beyond sublinear changes, we slightly adapt the previous procedure by replacing the critical value c_α by a threshold sequence $c_m \to \infty$, arbitrarily slowly as $m \to \infty$. The original procedure was designed to have an asymptotic false alarm probability of α in the in-control situation. For late changes with $\beta \geqslant 1$, this leads to a positive probability of an alarm before the change occurs, which contaminates the findings of the detection delay. Replacing c_α by $c_m \to \infty$ avoids this problem. For early changes with $\beta < 1$, this adaptation is not necessary because such changes occur in an asymptotic sense almost immediately after the monitoring starts.

A more detailed discussion can be found leading up to Proposition 1 of Kirch & Stoehr (2022a).

In this section we concentrate on the classical weight function (6) with $\gamma = 0$. For early changes with $\beta < 1$, the analysis has also been done for other values of $\gamma < 1/2$; see Theorem 5.3 in the 2019 PhD thesis by C. Stoehr from the Otto-von-Guericke University, Magdeburg as well as the 2003 Universität zu Köln dissertation by A. Aue, Aue & Horváth (2004) and Fremdt (2015). Delay times can also be computed for the sequential tests with $\gamma = 1/2$. Despite the extreme-value limit in this case, see (12), the delay times are asymptotically normal as shown in Aue et al. (2008). These changes correspond to the situation where the change occurs almost immediately after monitoring starts. In this situation a γ closer to $1/2$ can decrease the detection delay at the cost of a higher probability of a false alarm at the very beginning of the monitoring period; see Remark 4 of Kirch & Stoehr (2022a). However, for linear and superlinear changes, the simulations of Kirch & Weber (2018) suggest that a value of $\gamma = 0$ is preferable for all changes that do not occur almost immediately after monitoring commences.

By Proposition 1 of Kirch & Stoehr (2022a), it holds that, for any $c_m \to \infty$ and any $k^* = \lfloor \lambda m^\beta \rfloor$ as $m \to \infty$,

$$P\left(\sup_{1 \leqslant k \leqslant k^*} \frac{|\Psi(m,k)|}{m^{1/2}(1+k/m)} > c_m \right) \to 0.$$

This motivates us to consider the following delay time instead of the stopping time:

$$\kappa_m := \inf\{k > k^* : w(m,k)|\Gamma(m,k)| > c_m\}.$$

The following result is a special case of Theorem 2 and Corollary 3, respectively, of Kirch & Stoehr (2022a), where here additional assumptions are imposed on c_m to help clarity of exposition. Recall from Theorem 2 that $m^{1/2}|d_m| \to \infty$ is required to obtain asymptotic power one such that $m^{1/2}|d_m|$ can be thought of as the signal strength. In the following theorem a slightly stronger assumption is needed to allow for the increasing threshold c_m.

THEOREM 3. *Let model (5) hold with $\{\varepsilon_i : i \geqslant 1\}$ independent and identically distributed, $\mathsf{E}(\varepsilon_i) = 0$, $0 < \sigma^2 = \mathsf{Var}(\varepsilon_i) < \infty$ and $k^* = \lfloor \lambda m^\beta \rfloor$ for some $\lambda > 0, \beta \geqslant 0$. Additionally, let $|d_m| = O_P(1)$ and $c_m \to \infty$ with $m^{1/2}|d_m|/c_m \to \infty$. Then the following assertions hold.*

(a) *It holds that, for all x,*

$$P\left(\frac{\kappa_m - a_m(T_m)}{b_m(T_m)} \leqslant x \,\middle|\, T_m \right) \xrightarrow{\mathrm{P}} \Phi\left(\frac{x}{\sigma}\right),$$

where $T_m = (1/\sqrt{m}) \sum_{i=1}^m \varepsilon_i$ and, with $\mathrm{sign}(d_m)$ denoting the sign of d_m,

$$a_m(T_m) = \left(1 - \frac{c_m}{m^{1/2}|d_m|} + \frac{T_m \,\mathrm{sign}(d_m)}{m^{1/2}|d_m|}\right)^{-1} \left(k^* + \frac{c_m\sqrt{m}}{|d_m|}\right),$$

$$b_m(T_m) = \frac{\sqrt{a_m(T_m)}}{|d_m|}$$

and Φ is the distribution function of a standard normal distribution.

(b) *If c_m is chosen such that $c_m^2/\log\log m \to \infty$ then the assertion also holds almost surely,*

$$P\left(\frac{\kappa_m - a_m(T_m)}{b_m(T_m)} \leqslant x \;\middle|\; T_m\right) \to \Phi\left(\frac{x}{\sigma}\right).$$

An application of Lebesgue's dominated convergence theorem gives the following unconditional result:

$$\frac{\kappa_m - a_m(T_m)}{b_m(T_m)} \xrightarrow{\text{D}} N(0, \sigma^2).$$

This shows in particular that $a_m(T_m)$ can be considered the expected detection delay, while $b_m(T_m)$ corresponds to the variability. The expected detection delay is smaller for larger changes $|d_m|$ and larger for later changes k^*. The same dependence on d_m and k^* also holds for the variability that increases with increasing error variance. The asymptotic approximation as given by Theorem 3 coincides well with the simulated small-sample behaviour of the delay time, as illustrated in § 4.1 of Kirch & Stoehr (2022a).

It is possible to replace $b_m(T_m)$ by $b_m(0)$ in the above result, where T_m in the formula is replaced by 0. In the sublinear case $a_m(T_m)$ can be replaced by $a_m(0)$. However, this substitution does not work in general; see Theorem 6 of Kirch & Stoehr (2022a). This shows that there is a phase transition between early, sublinear, changes, where the distribution of the delay time does not depend on the training data asymptotically, while the distribution does depend on the training data for later, linear and superlinear, changes. Furthermore, in the sublinear case the multiplicative factor $\{1 - c_m/(m^{1/2}|d_m|)\}^{-1}$ in $a_m(0)$ can be replaced by 1, leading to an expected delay of $\tilde{a}_m = k^* + c_m m^{1/2}/|d_m|$, which is the term obtained by Fremdt (2015). Despite the asymptotic equivalence of both terms, the small-sample approximation provided by $a_m(0)$ is much better than that based on $\tilde{a}_m$, where the latter systematically underestimates the delay time in small samples; see Fig. 3 of Kirch & Stoehr (2022a). For very early changes as in Aue & Horváth (2004) with $\beta < 1/4$, it is also possible to use $\check{a}_m = c_m m^{1/2}/|d_m|$ at the cost of approximation quality in small samples.

Sketch proof of Theorem 3. We only give the high-level idea of the proof going back to Aue & Horváth (2004), noting that complexity arises when filling in the missing details. For a more in-depth explanation, see also § 3.3 of Kirch & Stoehr (2022a). First, for arbitrary x, a_m and b_m, search for sequences e_m and v_m such that, with $P^*(\cdot) = P(\cdot \mid T_m = t_m)$,

$$P^*\left(\sup_{k^* < k \leqslant a_m + x b_m} \frac{|\Psi(m,k)|}{m^{1/2}(1 + k/m)} \leqslant e_m + z\sqrt{v_m}\right) \xrightarrow{\text{D}} \Phi(z) \quad \text{for all } z \in \mathbb{R}.$$

Note that e_m and v_m can be seen as expectation and variances of the supterm. Naturally, these sequences depend on a_m, b_m and x. The next task then is to choose a_m and b_m in such a way that

$$\frac{c_m - e_m}{\sqrt{v_m}} \to \frac{-x}{\sigma}.$$

This implies that

$$P^*\left(\frac{\kappa_m - a_m}{b_m} \leqslant x\right) = 1 - P^*(\kappa_m > a_m + xb_m)$$
$$= 1 - P^*\left(\sup_{k^* < k \leqslant a_m + xb_m} \frac{|\Psi(m,k)|}{m^{1/2}(1+k/m)} \leqslant e_m + \frac{c_m - e_m}{\sqrt{v_m}}\sqrt{v_m}\right)$$
$$\xrightarrow{\text{D}} 1 - \Phi\left(\frac{-x}{\sigma}\right) = \Phi\left(\frac{x}{\sigma}\right),$$

which completes the proof. □

In more complex situations the asymptotic distribution can even exhibit phase transitions for different $\beta < 1$, i.e., different sublinear regimes, see Fremdt (2014, Theorem 2.1) or Kirch & Stoehr (2022a, (2.13)).

2.5. *Variations of the monitoring scheme*

In § 2.2 we not only introduced one monitoring scheme, but a class of weighted CUSUM detectors under mild assumptions on the weight functions. Two main differences were pointed out. First, the choice of γ related to Assumption 1(i) most notably in (6) puts more emphasis on early changes that leads to shorter detection delays for very early changes. The price for this is paid in the form of longer detection delays for changes that do not occur almost immediately after monitoring starts. Second, both closed-end, $\rho(t) = 0$ for all $t > N$, as well as open-end, $\limsup_{t\to\infty} \rho(t) > 0$, procedures are included in this class. Effectively, these choices and, more generally, the shape of the weight function determine how to distribute the available probability of a false rejection among the different time-points k of the monitoring period in a similar spirit as the 'spending sequence' in Gandy (2009) and Lan & DeMets (1983) does, guaranteeing that the familywise error rate is controlled.

However, the monitoring scheme $|\Psi(m,k)|$ itself has a strong influence on the detection delay. As already pointed out, the CUSUM scheme based on $|\Psi(m,k)|$ effectively runs a two-sample test on the training dataset and all observed random variables from the start of the monitoring period. While being relatively simple and having desirable computational properties, the power of the corresponding test will not be ideal except for relatively early changes, because, for later changes, the monitoring period contains many observations stemming from the null distribution. Thus, significantly more evidence needs to be collected before this difference becomes significant in comparison to a test that compares the mean of all after-change observations with either the mean from the training data or the mean based on all before-change observations. Obviously, the latter tests are not available because the time of change is unknown, but various modifications have been considered in the literature to take this into account.

Indeed, the following additional monitoring schemes $\Gamma(m,k)$ have been proposed, where, for completeness, $\Gamma_1(m,k) = |\Psi(m,k)|$ corresponds to the previously considered CUSUM scheme:

$$\Gamma_1(m,k) := \left|\sum_{i=m+1}^{m+k} (X_i - \bar{X}_m)\right| \qquad \text{(CUSUM)},$$

$$\Gamma_2(m,k) := \left| \sum_{i=m+\lfloor kh_2 \rfloor+1}^{m+k} (X_i - \bar{X}_m) \right| \qquad \text{(mMOSUM)},$$

$$\Gamma_3(m,k) := \sup_{0\leqslant \ell<k} \left| \sum_{i=m+\ell+1}^{m+k} (X_i - \bar{X}_m) \right| \qquad \text{(Page-CUSUM)},$$

$$\Gamma_4(m,k) := \sum_{i=m+k-h_4+1}^{m+k} (X_i - \bar{X}_m) \qquad \text{(MOSUM)},$$

$$\Gamma_{5,1}(m,k) := \sup_{0\leqslant \ell<k} \left| \sum_{i=m+\ell+1}^{m+k} X_i - \frac{k-\ell}{m+\ell} \sum_{j=1}^{m+\ell} X_j \right| \qquad \text{(full-CUSUM}_1\text{)},$$

$$\Gamma_{5,2}(m,k) := \sup_{0\leqslant \ell<k} \frac{m+\ell}{m} \left| \sum_{i=m+\ell+1}^{m+k} X_i - \frac{k-\ell}{m+\ell} \sum_{j=1}^{m+\ell} X_j \right| \qquad \text{(full-CUSUM}_2\text{)}.$$

Note that Γ_1 through $\Gamma_{5,2}$ are based on differences between a monitoring sample mean, computed from the latest set of observations in the monitoring period, and a reference sample mean. This reference sample mean can be the sample mean of the training data or the sample mean of all those observations not used to calculate the monitoring sample mean. Both moving sum, MOSUM, methods only take more recent observations from the monitoring period into account, where the number of these observations increases with time for Γ_2, but not for Γ_4. Schemes Γ_3, $\Gamma_{5,1}$ and $\Gamma_{5,2}$ take a supremum with respect to the starting points from which to calculate the corresponding monitoring sample means. In all cases, the idea is to decrease the detection delay for later changes.

While Γ_2 to Γ_4 are still based on a comparison of the training dataset with the monitoring dataset, the $\Gamma_{5,\cdot}$, full-CUSUMS, compare the sample mean of the data obtained after a point ℓ in the monitoring dataset with the sample mean calculated from the training data augmented with the monitoring data up to ℓ and then maximize with respect to ℓ. Indeed, the latter is very closely related to likelihood ratios and, in particular, to the CUSUM-type statistics typically used in a posteriori changepoint analysis. Both Γ_2, MMOSUM and Γ_4, MOSUM are based on moving sums, where $0 < h_2 < 1$ is a fixed tuning bandwidth parameter independent of m, while $h_4 = h_{4,m} \in \mathbb{N}$ depends on m, where $\lim_{m\to\infty} h_4/m \to b \in [0,1]$. The limit distributions for $b = 0$ and $0 < b \leqslant 1$ differ notably. In both cases, the tuning parameters determine how many past observations from the monitoring dataset are being used for a comparison. While for MMOSUM the number of observations from the monitoring dataset grows with k, namely, $k - \lfloor kh_2 \rfloor$, for fixed m, it remains fixed at h_4 for the MOSUM scheme. Page-CUSUM Γ_3 as well as the full-CUSUMS $\Gamma_{5,\cdot}$ do not rely on such tuning parameters, but instead take the supremum over all possible choices. Page-CUSUM always uses the training dataset as the first sample to compare with, while the full-CUSUMS split the full data into two sets using all available data for any choice of ℓ.

Kirch & Weber (2018) conducted a comparison of the monitoring schemes Γ_1 to Γ_4, also using different weight functions. A similar comparison of Γ_1 to Γ_3 was considered by Kirch & Stoehr (2022b). In summary, the detection power, i.e., the percentage of simulated time series with changes for which the procedure did stop, of the MMOSUM detector with

suitably chosen parameter h_2 is typically highest or at least very close to the highest. Generally, a choice of $h_2 = 0.4$ leads to very good power behaviour in all simulated situations. Page-CUSUM and CUSUM exhibit a similar detection power for the open-end procedure, but Page-CUSUM is typically superior for the closed-end procedure. The MOSUM statistic tends to have significantly lower detection power for the considered values of h_4, which does increase with increasing h_4. Dette & Gösmann (2020) gave a comparison of the detection power for CUSUM, Page-CUSUM and full-CUSUM$_1$, where the latter improves upon the former two. Additionally, they compared the average detection times. These turn out to be relatively similar to each other, as the outcome depends on the magnitude of the changepoint. Simulations comparing all five monitoring schemes based on the empirical density of the stopping times have yet to be added to the literature.

Generally, it is most illustrative to look at the empirical densities of the empirical stopping times, which are given by both Kirch & Weber (2018) and Kirch & Stoehr (2022b). These indicate that the detection delay is typically smallest for MMOSUM with suitable bandwidth, which should be smaller for earlier changes and larger for later changes. If MOSUM detects the change, it tends to do so quickly. Interestingly, CUSUM and Page-CUSUM seem to behave similarly in these simulations, but again MMOSUM with a choice of $h_2 = 0.4$ outperforms them, except for early changes where MMOSUM tends to be slightly slower.

Historically, Γ_1 was first introduced by Chu et al. (1996) and later analysed by Horváth et al. (2004). Stopping times for very early changes were derived by Aue & Horváth (2004). Then, Γ_4 was proposed by Zeileis et al. (2005), followed by a detailed asymptotic analysis by Horváth et al. (2008) and Aue et al. (2012a). Monitoring scheme Γ_2 was proposed by Chen & Tian (2010) and an error in their limit distribution was corrected by Kirch & Weber (2018). The original paper of Page (1954) motivated Fremdt (2015) to introduce Γ_3. The author further investigated stopping times for sublinear changes in Fremdt (2014). Recently, Dette & Gösmann (2020) and Gösmann et al. (2021) introduced $\Gamma_{5,1}$ for the closed- and open-end procedures, respectively, while Holmes & Kojadinovic (2021) introduced $\Gamma_{5,2}$. Most of these papers already discussed more general models such as linear models that include the above changes in expectation as a special case.

It would be of considerable interest for future work to derive the asymptotic distribution of the delay times for the different monitoring schemes and linear as well as superlinear changes, thus allowing a theoretical comparison of the delay times for change locations relevant for applications.

2.6. *Bootstrap procedures*

While the practitioner needs to decide on a weight function $w(m, \cdot)$ in advance, the corresponding critical value c_α is obtained from the asymptotic distribution in the no-change situation. The corresponding quantiles are explicitly known only for selected weight functions, for example (6) with $\gamma = 0$. In other situations, the critical value needs to be obtained by classical Monte Carlo simulations or by more refined methods such as those proposed by Franke et al. (2022). On the other hand, for a posteriori tests, it has been shown that resampling methods can often improve the small-sample behaviour of asymptotic tests. These can typically be implemented by standard Monte Carlo simulations or based on more refined methods, as discussed by Gandy (2009). Because we are dealing with an independent and identically distributed situation, under the null hypothesis, a standard bootstrap that samples independently with replacement from the original observations can be used. Theoretically, this means that the bootstrap replicates are an independent and identically

distributed sample from the empirical distribution function of the observations. While in the a posteriori situation, in which bootstrap methods are well understood, it is clear what the original observations are, this is much less clear in a sequential setting, where more and more observations are obtained until the procedure stops. The simplest idea is to use the training dataset for a standard, nonstudentized, bootstrap, which means replacing the critical value c_α with the smallest c_α^* fulfilling

$$P\Big(\sup_{k \geqslant 1} w(m,k)|\Psi^*(m,k)| > c_\alpha^* \sqrt{m} \;\Big|\; X_1, \ldots, X_m \Big) \leqslant \alpha,$$

where

$$\Psi^*(m,k) = \sum_{i=m+1}^{m+k} (X_{U_m(i)} - \bar{X}_m^*), \qquad \bar{X}_m^* = \frac{1}{m} \sum_{i=1}^{m} X_{U_m(i)}, \tag{13}$$

and $\{U_m(i) \colon i \geqslant 1\}$ is an independent and identically distributed sequence with $P(= \{U_m(i) = j\} = 1/m)$ for $j = 1, \ldots, m$. Numerically, the above conditional distribution can be approximated by Monte Carlo simulations, drawing bootstrap observations $X_{U_m(i)}$ by sampling with replacement from the training dataset. Typically, one gets better small-sample behaviour if studentized versions are used, which in this case means rejecting the null as soon as $w(m,k)|\Psi(m,k)| > c_\alpha^* m^{1/2} \hat{\sigma}_m$, with $\hat{\sigma}_m$ as in (8), and dividing $\Psi^*(m,k)$ in (13) by $\sigma_m^* = \{1/(m-1) \sum_{j=1}^m (\ldots)^2/(m-1)\}^{1/2}$. This bootstrap has the advantage that all original random variables come from the in-control situation: generally, a bootstrap scheme for testing should aim at mimicking the distribution under the null hypothesis to increase the small-sample power; see Hall & Wilson (1991).

On the other hand, a bootstrap scheme based on the training dataset only makes use of a small part of the data. Clearly, before the $(m+k)$th observation, we could recalculate the bootstrap critical values analogously to the above, but based on $X_1, \ldots, X_{m+k-1}$ rather than just the training dataset, i.e., with $\{U_m(\cdot)\}$ replaced by $\{U_{m,k}(\cdot)\}$ independent and identically distributed with $P\{U_m(i) = j\} = 1/(m+k-1)$ for $j = 1, \ldots, m+k-1$. This yields a sequence of critical values $c_\alpha^*(k)$ depending on k that increases the complexity of the theoretical analysis. From a computational perspective, it is expensive to completely recalculate the critical values after each new observation with little new information gain. Therefore, it is natural to only replace the oldest bootstrap samples by new ones obtained from all observations that have been observed up to that time-point and to do this only after a batch of new observations has been observed rather than after every single observation. Theoretically, this corresponds to drawing from a corresponding convex combination of the bootstrap distributions. While in this case, a certain percentage of the observations can be expected to be contaminated by a change as some observations are required before the procedure stops, typically, this percentage is relatively small, so that the associated power loss in turn tends to be small.

The above bootstrap methods were proposed by Kirch (2008), who also proved their theoretical validity in the sense that (7) still holds with the sequence of bootstrap critical values corresponding to the above procedures. Furthermore, extensive simulations confirmed that the bootstrap schemes increase the power, while having comparable size behaviour if more observations are being used in the procedure. Because the difference between a bootstrap based on all available observations and the proposed convex combination is relatively small, but the computational difference is sizeable, such a convex combination is highly preferable.

For increasing the length of the training dataset, all procedures become nearly equivalent, in line with their theoretical validity.

It has been observed in the literature that using weight functions (6) with γ closer to $1/2$ leads to a significant rate of false alarms at the very beginning of the training dataset. This motivated Kirch & Weber (2018) to only start monitoring after l_m observations of the monitoring data; see Remark 1. Interestingly, the above bootstrap procedures also seem to significantly reduce this effect.

3. Extensions to more complex data structures

3.1. *Regression models*

The presentation of this paper focuses on the location-shift model to improve readability. However, many of the papers cited above dealt with more complex data-generating mechanisms, one of the most prominent being the linear regression model. For example, the original paper of Chu et al. (1996) focused on the linear regression $Y_i = X_i^{\mathrm{T}}\beta_i + \varepsilon_i$, with random k-dimensional predictors $\{X_i\}$ satisfying regularity conditions and a martingale difference sequence $\{\varepsilon_i\}$. Chu et al. (1996) used different weight functions than those defined through (6) to calibrate the CUSUM detector. Their work was followed by Horváth et al. (2004), who provided test statistics based on (6) and another weight function and derived their asymptotic behaviour under both the null and alternative for errors satisfying strong invariance principles. Aue et al. (2006) introduced a new test statistic based on squared regression residuals in a regression setting with martingale difference errors, also focusing on testing. The corresponding stopping times were, for example, considered by Aue et al. (2009), who studied regression models for which the predictors may be specified as time series. A version of Page's CUSUM for time series regression was studied by Fremdt (2015). In Hušková & Kirch (2012) both a regression and a pair bootstrap for sequential tests in linear models were analysed theoretically and in simulations, where the bootstrap critical values are updated as more and more data become available following the approach of Kirch (2008). Recently, Horváth et al. (2022a) proposed a sequential monitoring procedure in linear models to monitor for changes in the real estate market. Zhou et al. (2015) proposed an open and closed-end procedure for quantile regression models. Nonlinear regression models were the subject of Ciuperca (2013).

3.2. *Time series errors*

In many applications the assumption of independent and identically distributed errors is too restrictive. However, a closer look at the proof methods reveals that results carry over to time series errors as long as these fulfil a corresponding functional central limit theorem in combination with the necessary Hájék–Rényi inequalities. Both are fulfilled under strong invariance principles that have been verified for a variety of nonparametric time series concepts, including mixing (Bradley, 2005; Doukhan, 2012), generalizations of m approximability (Wu, 2005) and others. In these situations, the asymptotic variance in the functional central limit theorem, and in extension the limit distribution of the corresponding monitoring procedure, is given by the long-run variance that takes the full covariance structure into account. Unfortunately, it is notoriously difficult to estimate with good small-sample properties, in particular in the presence of changepoints. In particular, this situation is included in the framework of Kirch & Tadjuidje Kamgaing (2015), Kirch & Weber (2018) and Kirch & Stoehr (2022a,b), as discussed in § 2 as well as many of the papers in § 3.1,

starting with the early works of Horváth et al. (2004) on testing and Aue & Horváth (2004) on delay times.

3.3. *General construction principle based on estimating functions*

Many of the above procedures for changes in the location or a linear model can be framed in the more general setting of estimating functions, also known as the generalized method of moments, which is strongly related to M-estimation. Such estimators $\hat{\theta}_m$ are obtained as solutions of the equation $\sum_{i=1}^{m} H(Z_i, \theta) = 0$, where H is a suitable estimating function and the Z_i are the observations, possibly including lagged observations in the case of autoregressive models. Monitoring schemes are typically based on $\Psi_H(m,k) = \sum_{i=m+1}^{m+k} H(X_i, \hat{\theta}_m)$, including several of the approaches mentioned above. For example, $H(z,\theta) = x - \theta$ corresponds to minimizing the least-squares loss and leads to the procedures for the location model discussed in §2 if applied to $Z_i = X_i$ and $\theta = \mu$, because in this case the corresponding estimator is given by $\hat{\mu}_m = \bar{X}_m$. For linear regression models with $Y_i = X_i^{\mathrm{T}}\beta + \varepsilon_i$, the parameter β can be estimated via least squares using $H(Z_i, \beta) = X_i(Y_i - X_i^{\mathrm{T}}\beta)$, $Z_i = (Y_i, X_i^{\mathrm{T}})^{\mathrm{T}}$. The first papers on sequential tests in linear regression models, such as Chu et al. (1996) and Horváth et al. (2004), constructed the monitoring scheme instead on $\tilde{H}(Z_i, \beta) = Y_i - X_i^{\mathrm{T}}\beta$, i.e., on a partial sum process of the estimated residuals using an estimator for β based only on the training dataset. Theory for this setting with $\Psi_{\tilde{H}}(m,k) = \sum_{i=m+1}^{m+k} \tilde{H}(X_i, \hat{\theta}_m)$ for a suitable $\tilde{H} \neq H$ is also included in the general theory discussed by Kirch & Tadjuidje Kamgaing (2015). However, in the linear regression context, such a procedure mainly has power against alternatives that go along with a change in mean, but has little power otherwise. In contrast, the full procedure based on the monitoring statistics with H in the notation of this section, as proposed by Hušková & Koubková (2005), has power against all changes in the regression parameter. In a time series context, often lagged observations are included in Z_i, for example, for a simple first-order autoregressive, AR(1), model with $X_t = aX_{t-1} + \varepsilon_t$, $t = 0, 1, 2, \ldots$, one would typically choose $Z_i = (X_i, X_{i-1})$, $i = 1, \ldots, m$, $\theta = a$ and $H(Z_i, a) = X_{i-1}(X_i - aX_{i-1})$, again corresponding to least-squares estimation also related to Yule–Walker estimators in this case. For nonlinear parametric models, it is natural to use likelihood theory that also fits in this framework via the score function. This leads, for example, to monitoring statistics previously considered for generalized autoregressive conditional heteroscedastic time series or count time series, discussed below. Similarly, by using more robust, often bounded, estimating functions, robust methods are also included in this framework; see §3.5 below. Theoretical discussions for several of the monitoring schemes of §2.5 in the framework of estimating functions can be found in Kirch & Tadjuidje Kamgaing (2015) and Kirch & Weber (2018). The theory of Kirch & Tadjuidje Kamgaing (2015) allows for the use of a different estimating function for the initial estimation of $\hat{\theta}_m$ and for the monitoring scheme $\Psi_{\tilde{H}}$, as mentioned above. Such an approach can be advantageous if the training dataset does not contain any outliers, while the new incoming data are expected to contain outliers.

3.4. *Changes in the time series structure and other stochastic processes*

Statistical analysis of dependent data is often based on parametric time series models such as autoregressive moving average, ARMA, processes in the linear case and autoregressive- or generalized-autoregressive-conditional-heteroscedasticity-, ARCH/GARCH, type processes in the nonlinear case. In certain situations it is of interest to monitor the parameters of an initially fixed model for continued appropriateness. In order to help with this goal, the sequential procedures discussed in §2 were extended to cover online tests

for the constancy of the parameters in many of the popular time series models. The case of autoregressive processes is treated by Gombay & Serban (2009). Hlávka et al. (2016) discussed bootstrap versions of sequential tests for autoregressive processes. Hlávka et al. (2012) analysed both closed-end and open-end sequential changepoint tests for changes in the error distribution of autoregressive time series where the monitoring statistic is based on the empirical characteristic function. The procedure does not only have power with respect to changes in the error distribution, but also with respect to changes in the autoregressive parameters. Because the limit distribution contains unknown parameters that are difficult if not impossible to estimate, a bootstrap based only on the training dataset is also proposed and its theoretical validity shown. More complicated bootstrap procedures where the critical values are updated with new incoming monitoring data are empirically investigated by Hlávka et al. (2016), where a power comparison with sequential Kolmogorov–Smirnov-type statistics that is based on the empirical distribution is also given rather than the characteristic function. Hlávka et al. (2020) proposed both a sequential closed-end procedure and an a posteriori procedure based on empirical characteristic functions for testing the null hypothesis of two jointly observed time series having the same marginal distribution against the alternative that at some unknown point in time the marginal distributions become different. Their sequential setting is not based on the notion of a training dataset except for their proposed bootstrap procedure. Online tests for more general ARMA processes were treated by Dienes & Aue (2014), while the corresponding delay times were derived by Aue et al. (2015). Hoga (2017) and Gösmann et al. (2021) introduced sequential tests for multivariate time series, while Hoga & Wied (2017) discussed sequential monitoring procedures for the tail behavior of β-mixing time series. Kurozumi (2020, 2021, 2023) proposed a sequential procedure for the detection of bubbles, i.e., the situation where a unit-root process changes to an explosive process. Sequential tests for time series behaviour switching from stationary to mildly nonstationary can be found in Horváth et al. (2020).

Turning to nonlinear time series models, Berkes et al. (2004) suggested monitoring schemes based on the loglikelihood score function to monitor for changes in GARCH(p, q) models. Online tests for constant variance in heteroscedastic time series were proposed by Horváth et al. (2006). Sequential tests for conditional correlation models were treated by Pape et al. (2021). Na et al. (2010) and Horvath & Trapani (2021) investigated the properties of online monitoring procedures for random coefficient autoregressive time series. Furthermore, Kirch & Tadjuidje Kamgaing (2015) proposed to use least-squares methods for neuronal networks to detect changes in nonparametric nonlinear autoregressive models with continuous errors. Na et al. (2011) proposed a monitoring procedure for detection of parameter changes in parametric time series models directly based on the difference between parameter estimators from the training dataset and all observations under certain assumptions on the regularity of these estimators. A similar approach was taken by Bardet & Kengne (2014) for general causal time series, including AR(∞), ARCH(∞), threshold autoregressive conditionally heteroscedastic (TARCH)(∞) or ARMA-GARCH models.

Particularly challenging are count time series, where the observations are discrete and serially correlated. Kirch & Tadjuidje Kamgaing (2015) and Kirch & Weber (2018) proposed and analysed sequential monitoring tests for binary autoregressive and Poisson autoregressive time series. Their methods can also be dealt with in the setting of estimating functions, as outlined in § 3.3. For integer-valued autoregressive conditionally heteroscedastic (INARCH) time series, Hudecová et al. (2016) proposed and analysed sequential tests based on the empirical probability generating function. A similar approach is taken by

Hudecová et al. (2017) for more general time series of counts, including a corresponding bootstrap based only on the training dataset.

Monitoring procedures for renewal processes were analysed by Gut & Steinebach (2002), where extreme-value asymptotics are derived for a closed-end procedure without a training dataset. Asymptotic results for the corresponding stopping times for early changes were derived by Gut & Steinebach (2009). Mihalache (2012) proposed a procedure to detect changes in diffusion processes. Steinebach (2009) proposed a monitoring procedure based on a training dataset for changes in the adjustment coefficient of a ruin model perturbed by diffusion. Finally, Timmermann (2015) analysed the behaviour of sequential procedures for gradual changes in the location of a general stochastic process, including renewal processes; see also the 2014 University of Cologne PhD thesis by H. Timmermann.

3.5. *Robust methods*

The methodology introduced in the previous sections is based on sample means. Its asymptotic theory relies heavily on the existence of at least two moments and is therefore not suitable for very heavy-tailed data, with infinite variance, or in the presence of outliers. Indeed, this is a classical setting both for using tests based on M-estimation and for methodology based on U-statistics extending Mann–Whitney–Wilcoxon statistics. Such methodology can also be extended to sequential changepoint testing. In particular, the methodology based on estimating functions as detailed by Kirch & Tadjuidje Kamgaing (2015) and Kirch & Weber (2018) allows the use of more robust methodology for location changes; see § 3.3 for more details. However, their regularity assumptions exclude classical nondifferentiable estimators such as the signum function corresponding to the median. The latter is discussed in detail for independent and identically distributed errors in the 2006 PhD thesis of A. Koubková, and was extended to α-mixing time series errors in the 2013 PhD thesis of O. Chochola, both from Charles University. A different approach is taken by Kirch & Stoehr (2022a,b), who considered general monitoring schemes based on U-statistics with time series errors. This includes the classical location model as in (5), but also more importantly a sequential monitoring scheme related to the Mann–Whitney–Wilcoxon statistic. It is shown in simulations and a theoretical comparison of the delay times that the latter outperforms the former for heavy-tailed distributions.

3.6. *General distributional changes*

Hlávka et al. (2012, 2016) considered linear autoregressive time series models, but then constructed monitoring procedures based on the empirical characteristic function of estimated residuals. As such, the corresponding test statistics do not only have power against changes in the parameters of the linear time series, but also against changes in the error distribution. Furthermore, bootstrap procedures are discussed for this statistic. While their monitoring scheme is clearly much more complicated than Ψ in § 2, they only considered the corresponding classical CUSUM scheme. More recently, Kojadinovic & Verdier (2021) and Holmes et al. (2022) proposed a procedure for the detection of distributional changes in a multivariate version of the location model, as in (5). Their procedure is not based on the empirical characteristic, but on the empirical distribution function and they considered the more recent CUSUM scheme $\Gamma_{5,2}$. In addition, Horváth et al. (2021) discussed the monitoring of a sequence of distributions with a detector built from a weighted Wasserstein distance.

3.7. *Changes in functional, high-dimensional and other complex data*

Sequential methods for complex data, such as functions, high-dimensional vectors and networks, comprise a currently active subarea of research, with many significant papers still being published. We summarize here some of the results related to this growing body of work. A functional data set-up was considered by Aue et al. (2014), who discussed sequential monitoring procedures for dependent functional linear models with the goal of monitoring daily pollution curves. An online procedure to test for the stability of high-frequency portfolio betas in a capital asset pricing model was considered by Aue et al. (2012b) and in the 2013 Charles University PhD thesis of O. Chochola. Functional event data such as financial data related to initial public offerings of stock was studied by Horváth et al. (2022b).

High-dimensional settings were considered in an approach slightly different from that primarily discussed here in Liu et al. (2021) and Chen et al. (2022, 2023). High-frequency data were the subject of Romano et al. (2023), who applied their functional pruning CUSUM statistics to computer server data. All the previous four papers emphasize the need for the sequential algorithms to circumvent computational constraints. Another paper in the high-dimensional setting is that of Gösmann et al. (2022), who considered sequential changepoint detection in high-dimensional time series.

Other very recent papers on complex data include the paper by Berrett & Yu (2021), who studied online changepoint detection problems under the constraint of local differential privacy where the statistician does not have access to the raw data. As a concrete problem, the authors studied a multivariate nonparametric regression problem with the primary aim of detecting changes in the regression function. Online network changepoint methods were introduced by Yu et al. (2021), where networks are of Bernoulli type. Dubey et al. (2023) and Xu et al. (2023) considered online network changepoint detection with missing values, with the latter paper dealing with the dependent case. Graph-based online procedures for high-dimensional and non-Euclidean data were discussed by Chu & Chen (2022). He et al. (2021) developed online procedures for matrix-valued time series.

References

Aston, J. A. & Kirch, C. (2018). High dimensional efficiency with applications to change point tests. *Electron. J. Statist.* **12**, 1901–47.

Aue, A., Dienes, C., Fremdt, S. & Steinebach, J. (2015). Reaction times of monitoring schemes for ARMA time series. *Bernoulli* **21**, 1238–59.

Aue, A., Hörmann, S., Horváth, L. & Hušková, M. (2014). Dependent functional linear models with applications to monitoring structural change. *Statist. Sinica* **24**, 1043–73.

Aue, A., Hörmann, S., Horváth, L., Hušková, M. & Steinebach, J. G. (2012b). Sequential testing for the stability of high-frequency portfolio betas. *Economet. Theory* **28**, 804–37.

Aue, A. & Horváth, L. (2004). Delay time in sequential detection of change. *Statist. Prob. Lett.* **67**, 221–31.

Aue, A. & Horváth, L. (2013). Structural breaks in time series. *J. Time Ser. Anal.* **34**, 1–16.

Aue, A., Horváth, L., Hušková, M. & Kokoszka, P. (2006). Change-point monitoring in linear models. *Economet. J.* **9**, 373–403.

Aue, A., Horváth, L., Kokoszka, P. & Steinebach, J. (2008). Monitoring shifts in mean: Asymptotic normality of stopping times. *Test* **17**, 515–30.

Aue, A., Horváth, L., Kühn, M. & Steinebach, J. (2012a). On the reaction time of moving sum detectors. *J. Statist. Plan. Infer.* **142**, 2271–2288.

Aue, A., Horváth, L. & Reimherr, M. (2009). Delay times of sequential procedures for multiple time series regression models. *J. Economet.* **149**, 174–90.

Aue, A. & Kühn, M. (2008). Extreme value distribution of a recursive-type detector in a linear model. *Extremes* **11**, 135–66.

Bardet, J.-M. & Kengne, W. (2014). Monitoring procedure for parameter change in causal time series. *J. Mult. Anal.* **125**, 204–21.

Benincasa, G., Mansueto, G. & Napoli, C. (2019). Fluid-based assays and precision medicine of cardiovascular diseases: the 'hope' for Pandora's box? *J. Clin. Pathol.* **72**, 785–99.
Berkes, I., Gombay, E., Horváth, L. & Kokoszka, P. (2004). Sequential change-point detection in GARCH(p, q) models. *Economet. Theory* **20**, 1140–67.
Berrett, T. & Yu, Y. (2021). Locally private online change point detection. In *Advances in Neural Information Processing Systems*, vol. 34, Ed. M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang and J. Wortman Vaughan, pp. 3425–37. Red Hook, NY: Curran Associates.
Bradley, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Prob. Surveys* **2**, 107–44.
Chen, Y., Wang, T. & Samworth, R. J. (2022). High-dimensional, multiscale online changepoint detection. *J. R. Statist. Soc.* B **84**, 234–66.
Chen, Y., Wang, T. & Samworth, R. J. (2023). Inference in high-dimensional online changepoint detection. *J. Am. Statist. Assoc.*, doi: 10.1080/01621459.2023.2199962.
Chen, Z. & Tian, Z. (2010). Modified procedures for change point monitoring in linear models. *Math. Comput. Simul.* **81**, 62–75.
Cho, H. & Kirch, C. (2021). Data segmentation algorithms: univariate mean change and beyond. *Economet. Statist.*, doi: 10.1016/j.ecosta.2021.10.008.
Chu, C.-S., Stinchcombe, M. & White, H. (1996). Monitoring structural change. *Econometrica* **64**, 1045–65.
Chu, L. & Chen, H. (2022). Sequential change-point detection for high-dimensional and non-Euclidean data. *IEEE Trans. Sig. Proces.* **70**, 4498–511.
Ciuperca, G. (2013). Two tests for sequential detection of a change-point in a nonlinear model. *J. Statist. Plan. Infer.* **143**, 1719–43.
Csörgő, M. & Horváth, L. (1997). *Limit Theorems in Change-Point Analysis*. New York: John Wiley.
Dette, H. & Gösmann, J. (2020). A likelihood ratio approach to sequential change point detection for a general class of parameters. *J. Am. Statist. Assoc.* **115**, 1361–77.
Dienes, C. & Aue, A. (2014). On-line monitoring of pollution concentrations with autoregressive moving average time series. *J. Time Ser. Anal.* **35**, 239–61.
Doukhan, P. (2012). *Mixing: Properties and Examples* (Lecture Notes Statist. **85**). New York: Springer.
Dubey, P., Xu, H. & Yu, Y. (2023). Online network change point detection with missing values and temporal dependence. *arXiv:* 2110.06450v2.
Franke, J., Hefter, M., Herzwurm, A., Ritter, K. & Schwaar, S. (2022). Adaptive quantile computation for brownian bridge in change-point analysis. *Comp. Statist. Data Anal.* **167**, 107375.
Fremdt, S. (2014). Asymptotic distribution of the delay time in Page's sequential procedure. *J. Statist. Plan. Infer.* **145**, 74–91.
Fremdt, S. (2015). Page's sequential procedure for change-point detection in time series regression. *Statistics* **49**, 128–55.
Gandy, A. (2009). Sequential implementation of Monte Carlo tests with uniformly bounded resampling risk. *J. Am. Statist. Assoc.* **104**, 1504–11.
Gombay, E. & Serban, D. (2009). Monitoring parameter change in AR(p) time series models. *J. Mult. Anal.* **100**, 715–25.
Gösmann, J., Kley, T. & Dette, H. (2021). A new approach for open-end sequential change point monitoring. *J. Time Ser. Anal.* **42**, 63–84.
Gösmann, J., Stoehr, C., Heiny, J. & Dette, H. (2022). Sequential change point detection in high dimensional time series. *Electron. J. Statist.* **16**, 3608–71.
Gut, A. & Steinebach, J. (2002). Truncated sequential change-point detection based on renewal counting processes. *Scand. J. Statist.* **29**, 693–719.
Gut, A. & Steinebach, J. (2009). Truncated sequential change-point detection based on renewal counting processes II. *J. Statist. Plan. Infer.* **139**, 1921–36.
Hall, P. & Wilson, S. R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics* **47**, 757–62.
He, Y., Kong, X., Trapani, L. & Yu, L. (2021). Online change-point detection for matrix-valued time series with latent two-way factor structure. *arXiv:* 2112.13479v1.
Hlávka, Z., Hušková, M., Kirch, C. & Meintanis, S. G. (2012). Monitoring changes in the error distribution of autoregressive models based on Fourier methods. *Test* **21**, 605–34.
Hlávka, Z., Hušková, M., Kirch, C. & Meintanis, S. G. (2016). Bootstrap procedures for online monitoring of changes in autoregressive models. *Commun. Statist.* B **45**, 2471–90.
Hlávka, Z., Hušková, M. & Meintanis, S. G. (2020). Change-point methods for multivariate time-series: paired vectorial observations. *Statist. Papers* **61**, 1351–83.
Hoga, Y. (2017). Monitoring multivariate time series. *J. Mult. Anal.* **155**, 105–21.
Hoga, Y. & Wied, D. (2017). Sequential monitoring of the tail behavior of dependent data. *J. Statist. Plan. Infer.* **182**, 29–49.
Holmes, M. & Kojadinovic, I. (2021). Open-end nonparametric sequential change-point detection based on the retrospective CUSUM statistic. *Electron. J. Statist.* **15**, 2288–335.

Holmes, M., Kojadinovic, I. & Verhoijsen, A. (2022). Multi-purpose open-end monitoring procedures for multivariate observations based on the empirical distribution function. *J. Time Ser. Anal.* **45**, 27–56.
Horváth, L., Hušková, M., Kokoszka, P. & Steinebach, J. (2004). Monitoring changes in linear models. *J. Statist. Plan. Infer.* **126**, 225–51.
Horváth, L., Kokoszka, P. & Steinebach, J. (2007). On sequential detection of parameter changes in linear regression. *Statist. Prob. Lett.* **77**, 885–95.
Horváth, L., Kokoszka, P. & Wang, S. (2021). Monitoring for a change point in a sequence of distributions. *Ann. Statist.* **49**, 2271–91.
Horváth, L., Kokoszka, P. & Zhang, A. (2006). Monitoring constancy of variance in conditionally heteroskedastic time series. *Economet. Theory* **22**, 373–402.
Horváth, L., Kühn, M. & Steinebach, J. (2008). On the performance of the fluctuation test for structural change. *Seq. Anal.* **27**, 126–40.
Horváth, L., Liu, Z. & Lu, S. (2022a). Sequential monitoring of changes in dynamic linear models, applied to the US housing market. *Economet. Theory* **38**, 209–72.
Horváth, L., Liu, Z., Rice, G. & Wang, S. (2020). Sequential monitoring for changes from stationarity to mild non-stationarity. *J. Economet.* **215**, 209–38.
Horváth, L., Liu, Z., Rice, G., Wang, S. & Zhan, Y. (2022b). Testing stability in functional event observations with an application to IPO performance. *J. Bus. Econ. Statist.* **41**, 1–12.
Horváth, L. & Trapani, L. (2021). Changepoint detection in random coefficient autoregressive models. *arXiv:* 2104.13440v1.
Hudecová, Š., Hušková, M. & Meintanis, S. (2016). Change detection in INARCH time series of counts. In *Nonparametric Statistics*, Ed. R. Cao, W. González Manteiga and J. Romo, pp. 47–58. Cham: Springer.
Hudecová, Š., Hušková, M. & Meintanis, S. G. (2017). Tests for structural changes in time series of counts. *Scand. J. Statist.* **44**, 843–65.
Hušková, M. & Kirch, C. (2012). Bootstrapping sequential change-point tests for linear regression. *Metrika* **75**, 673–708.
Hušková, M. & Koubková, A. (2005). Monitoring jump changes in linear models. *J. Statist. Res.* **39**, 51–70.
Kirch, C. (2008). Bootstrapping sequential change-point tests. *Seq. Anal.* **27**, 330–49.
Kirch, C. & Stoehr, C. (2022a). Asymptotic delay times of sequential tests based on U-statistics for early and late change points. *J. Statist. Plan. Infer.* **221**, 114–35.
Kirch, C. & Stoehr, C. (2022b). Sequential change point tests based on U-statistics. *Scand. J. Statist.* **49**, 1184–214.
Kirch, C. & Tadjuidje Kamgaing, J. (2015). On the use of estimating functions in monitoring time series for change points. *J. Statist. Plan. Infer.* **161**, 25–49.
Kirch, C. & Weber, S. (2018). Modified sequential change point procedures based on estimating functions. *Electron. J. Statist.* **12**, 1579–613.
Kojadinovic, I. & Verdier, G. (2021). Nonparametric sequential change-point detection for multivariate time series based on empirical distribution functions. *Electron. J. Statist.* **15**, 773–829.
Krawczyk, B., Minku, L. L., Gama, J., Stefanowski, J. & Woźniak, M. (2017). Ensemble learning for data stream analysis: a survey. *Info. Fusion* **37**, 132–56.
Kumar, H. & Purbey, S. (2021). Green internet of things (IOT): a survey and future prospect. *Int. Res. J. Engineering Technol.* **8**, 1947–55.
Kurozumi, E. (2020). Asymptotic properties of bubble monitoring tests. *Economet. Rev.* **39**, 510–38.
Kurozumi, E. (2021). Asymptotic behavior of delay times of bubble monitoring tests. *J. Time Ser. Anal.* **42**, 314–37.
Kurozumi, E. (2023). Fluctuation-type monitoring test for explosive behavior. *Economet. Statist.*, doi: 10.1016/j.ecosta.2023.06.007.
Lan, K. K. G. & DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**, 659–63.
Lin, Z. & Bai, Z. (2010). *Probability Inequalities*. Berlin: Springer.
Liu, H., Gao, C. & Samworth, R. J. (2021). Minimax rates in sparse, high-dimensional change point detection. *Ann. Statist.* **49**, 1081–112.
Lorden, G. (1971). Procedures for reacting to a change in distribution. *Ann. Math. Statist.* **42**, 1897–908.
Mihalache, S. (2012). Strong approximations and sequential change-point analysis for diffusion processes. *Statist. Prob. Lett.* **82**, 464–72.
Milella, A., Reina, G. & Nielsen, M. (2019). A multi-sensor robotic platform for ground mapping and estimation beyond the visible spectrum. *Precis. Agric.* **20**, 423–44.
Moustakides, G. V. (1986). Optimal stopping times for detecting changes in distributions. *Ann. Statist.* **14**, 1379–87.
Na, O., Lee, J. & Lee, S. (2010). Monitoring parameter changes for random coefficient autoregressive models. *J. Korean Statist. Soc.* **39**, 281–8.
Na, O., Lee, Y. & Lee, S. (2011). Monitoring parameter change in time series models. *Statist. Meth. Applic.* **20**, 171–99.

Olsson, P.-O., Lindström, J. & Eklundh, L. (2016). Near real-time monitoring of insect induced defoliation in subalpine birch forests with MODIS derived NDVI. *Remote Sens. Environ.* **181**, 42–53.
Page, E. (1954). Continuous inspection schemes. *Biometrika* **41**, 100–15.
Pape, K., Galeano, P. & Wied, D. (2021). Sequential detection of parameter changes in dynamic conditional correlation models. *Appl. Stoch. Mod. Bus. Industry* **37**, 475–95.
Polunchenko, A. S. & Tartakovsky, A. G. (2012). State-of-the-art in sequential change-point detection. *Methodol. Comp. Appl. Prob.* **14**, 649–84.
Ritov, Y. (1990). Decision theoretic optimality of the CUSUM procedure. *Ann. Statist.* **18**, 1464–9.
Roberts, S. (1966). A comparison of some control chart procedures. *Technometrics* **8**, 411–30.
Roberts, S. (2000). Control chart tests based on geometric moving averages. *Technometrics* **42**, 97–101.
Romano, G., Eckley, I. A., Fearnhead, P. & Rigaill, G. (2023). Fast online changepoint detection via functional pruning CUSUM statistics. *J. Mach. Learn. Res.* **24**, 1–36.
Shewhart, W. (1931). *The Economic Control of Quality of a Manufactured Product*. New York: Van Nostrand.
Shiryaev, A. N. (1961). The problem of the most rapid detection of a disturbance in a stationary process. *Soviet Math. Dokl*, **2**, 795–9.
Shiryaev, A. N. (2010). Quickest detection problems: fifty years later. *Seq. Anal.* **29**, 345–85.
Steinebach, J. G. (2009). Monitoring risk in a ruin model perturbed by diffusion. *Metrika* **70**, 205–24.
Tartakovsky, A. G. (2019). *Sequential Change Detection and Hypothesis Testing*. Boca Raton, FL: CRC Press.
Tartakovsky, A., Nikiforov, I. & Basseville, M. (2014). *Sequential Analysis: Hypothesis Testing and Change-point Detection*. Boca Raton, FL: Chapman and Hall/CRC.
Timmermann, H. (2015). Sequential detection of gradual changes in the location of a general stochastic process. *Statist. Prob. Lett.* **99**, 85–93.
Truong, C., Oudre, L. & Vayatis, N. (2020). Selective review of offline change point detection methods. *Sig. Proces.* **167**, 107299.
Wald, A. (1947). *Sequential Analysis*. New York: John Wiley.
Wu, W. B. (2005). Nonlinear system theory: another look at dependence. *Proc. Nat. Acad. Sci.* **102**, 14150–4.
Xie, L., Zou, S., Xie, Y. & Veeravalli, V. V. (2021). Sequential (quickest) change detection: classical results and new directions. *IEEE J. Sel. Areas Info. Theory* **2**, 494–514.
Xu, H., Dubey, P. & Yu, Y. (2023). Online network change point detection with missing values and temporal dependence. *arXiv:* 2110.06450v2.
Yu, Y., Padilla, O. H. M., Wang, D. & Rinaldo, A. (2021). Optimal network online change point localisation. *arXiv:* 2101.05477v1.
Zeileis, A., Leisch, F., Kleiber, C. & Hornik, K. (2005). Monitoring structural change in dynamic econometric models. *J. Appl. Economet.* **20**, 99–121.
Zhou, M., Wang, H. J. & Tang, Y. (2015). Sequential change point detection in linear quantile regression models. *Statist. Prob. Lett.* **100**, 98–103.

[*Received on* 16 *October* 2023. *Editorial decision on* 12 *December* 2023]