On the use of estimating functions in monitoring time series for change points

Claudia Kirch[*] Joseph Tadjuidje Kamgaing[†]

January 12, 2015

Abstract

A large class of estimators including maximum likelihood, least squares and M-estimators are based on estimating functions. In sequential change point detection related monitoring functions can be used to monitor new incoming observations based on an initial estimator, which is computationally efficient because possible numeric optimization is restricted to the initial estimation. In this work, we give general regularity conditions under which we derive the asymptotic null behavior of the corresponding tests in addition to their behavior under alternatives, where conditions become particularly simple for sufficiently smooth estimating and monitoring functions. These regularity conditions unify and extend a large amount of existing procedures in the literature, while they also allow us to derive monitoring schemes in time series that have not yet been considered in the literature including non-linear autoregressive time series and certain count time series such as binary or Poisson autoregressive models. We do not assume that the estimating and monitoring function are equal or even of the same dimension, therefore, we allow for example to combine a non-robust but more precise initial estimator with a robust monitoring scheme. Some simulations and data examples illustrate the usefulness of the described procedures.

Keywords: Change analysis, nonparametric regression, nonlinear regression, autoregressive time series, sequential test, integer-valued time series

AMS Subject Classification 2000: 62L10,62G10,62M45

[*] Otto von Guericke University Magdeburg, Institute for Mathematical Stochastics, Postfach 4120 D–39016 Magdeburg, Germany; claudia.kirch@ovgu.de

[†] University Kaiserslautern, Department of Mathematics, Erwin-Schrödinger-Straße, D–67 653 Kaiserslautern, Germany; tadjuidj@mathematik.uni-kl.de

1 Introduction

In recent years, many data sets are collected automatically point-by-point, where a statistical decision needs to be reached online before the full data set is available. Examples include financial data sets, e.g., in risk management (Andreou and Ghysels [2]) or CAPM models (Aue et al. [5]) as well as medical data sets, e.g., monitoring intensive care patients (Fried and Imhoff [19]). With each new observation the question arises whether the model is still capable of explaining the data. If this is not the case an alarm needs to be raised, for example the financial models might not be anymore appropriate or the condition of the patient in intensive medical care might have changed. The statistical analysis of such data sets requires sequential methods, which are often called online monitoring procedures in the context of change point detection. Classical sequential change point procedures are often fully parametric and typically minimize the detection delay (as e.g. measured by the average delay time until an alarm is raised) under some restrictions on the false alarm rate (such as e.g. the average time until a false alarm is raised). However, such tests typically yield a false alarm with probability one, see, e.g., Anderson [1] or Siegmund [44].

More recently, Chu et al. [11] propose a different approach for monitoring structural changes in linear regression models. Their method controls the asymptotic type-I-error if no change occurs and has asymptotic power one under alternatives. Their key assumption is the existence of a historic data set without change which is used to estimate the unknown regression parameters before the change. In applications such a data set always exists as at least some data need to be collected before any reasonable statistical inference can be carried out. New incoming observations are then monitored for evidence of changes in these parameters. This framework allows for a rigorous asymptotic theory even for a possibly infinite monitoring horizon by letting the length of this historic data set grow to infinity without the need of any additional parametric assumptions on the error distribution. This framework has subsequently been extended in many different ways including detection of changes in non-linear time series such as GARCH models or more robust statistical methods (confer Section 6 for these and many more examples). While the corresponding sequential tests are based on very different statistics, all of them are constructed in a similar way and even the derivation of the asymptotic distribution shows many parallels.

In this paper, we explain this general construction principle and identify regularity conditions under which the asymptotic distribution of the corresponding tests can be obtained. This allows for the derivation of new monitoring procedures that have not been considered in the literature before such as monitoring procedures for integer-valued time series or a combination of robust and non-robust methods. The key to this unified theory are estimating functions (leading to estimating equations), which allow for a very general method of obtaining estimators including all classical examples (such as likelihood ratio estimators, method of moments estimators, least-squares estimator or M-estimators). Estimating functions are also known in the literature as objective functions or generalized method of moments. In the context of sequential monitoring procedures they are used in two different ways, first to get an initial estimator based on the historic data set and secondly to construct a score type detector statistic. The score type detector has the advantage that no sequential estimators are necessary which may cause numerical and computational problems in many situations.

In Section 2 we explain the general construction principles behind this class of sequential monitoring procedures. In Sections 3 and 4 we give regularity conditions under which we can derive the asymptotic distribution under the null hypothesis as well as the asymptotic power under alternatives. This is very general and does not require any particular parametric setup. For sufficiently smooth estimating and monitoring functions these regularity conditions essentially reduce to certain moment conditions and an appropriate weakly dependent property of the underlying process as illustrated in Section 5. In Section 6 we illustrate the usefulness of the monitoring procedure using nonlinear autoregressive count time series and the classical mean change model among others. For the existing examples, we include an intensive literature review, while new examples are accompanied by both a simulation study and data analysis. Finally, the proofs are given in Section 7.

2 Sequential testing based on estimating functions

In this section, we explain the general construction principles of sequential change-point tests in the spirit of Chu et al. [11]. For readers, who are not familiar with this methodology it may be helpful to keep the following two examples in mind:

mean change model: $X_t = \mu + e_t,$

where $\{e_t\}$ is a short-range dependent time series, or

non-linear autoregressive time series of order 1: $X_t = f_\theta(X_{t-1}) + e_t,$

where $\{f_\theta : \theta \in \Theta\}$ is a suitable class of functions leading to well defined stationary time series. The residuals $\{e_t\}$ are usually assumed to be i.i.d. in the literature but the general framework below also allows for dependent errors which arise for example if the monitoring is applied to data that do not follow the assumed parametric model. In particular this allows for a rigorous asymptotic treatment of the misspecified situation, in which changes are detected if the best approximating parameters before and after the change points differ. Similar ideas in the offline setup have been considered by Kirch and Tadjuidje Kamgaing [30].

We are now ready to explain how monitoring procedures can be derived in a general situation: The key assumption in Chu et al. [11], which we adopt here, states that there exists some historic data set which is stationary and does not contain a change. This so called non-contamination assumption is usually fulfilled in applications as some data need to be collected before any reasonable statistical inference can take place. Because this data set is stationary it can be used to estimate the initial set of parameters before any change occurs in a consistent way. Furthermore, the asymptotic considerations in the next sections will be carried out with respect to the length of this set to increase to infinity. To allow for a very general procedure, we only assume that this estimation is carried out with the use of an estimating function sometimes also called objective function or generalized method of moments. This is a very general framework in statistics to derive estimators including all classical examples such as maximum likelihood estimators, (weighted) least-squares estimators, M-estimators or moment estimators. The key is that the estimator is obtained as the solution of the following system of equations:

$$\sum_{t=1}^{m} G\left(\mathbf{X}_t, \widehat{\theta}_m\right) = 0, \tag{2.1}$$

where $\mathbf{X}_t, t = 1, \ldots, m$, are the historic observations and G (the so-called **estimating function**) is a suitable function with values in $\mathbb{R}^d$, where d is the number of unknown parameters in the parametric representation of interest. Note, that the left hand side is a vector so that the equality sign here (and elsewhere in the paper) is meant componentwise. Consequently, (2.1) is not an equation but a system of equations. The observations $\mathbf{X}_t$ are allowed to be multivariate, which is not only of importance in a truly multivariate setup but also e.g. in a regression situation with exogenous variables or even for an autoregressive setup of order p, where typically $\mathbf{X}_t$ consists of the past p elements of the autoregressive time series (confer Section 6 below). For reasonable estimating functions and under suitable regularity conditions, these estimators are consistent (as $m \to \infty$) for the parameter θ_0 defined by $\mathbb{E}G(\mathbf{X}_1, \theta_0) = 0$. In the correctly specified case this is the true parameter for reasonable estimating functions, while it is the best approximating parameter under misspecification.

New incoming observations $\mathbf{X}_{m+1}, \mathbf{X}_{m+2}, \ldots$ are monitored for a change in these estimated parameters. To this end, a score type detector statistic is used in order to avoid having to sequentially (re-)estimate the unknown parameters which might cause numerical and computational problems as the solution of (2.1) can often only be obtained numerically. An analogous representation to (2.1) is the key to this detector where we now consider the sum over the new observations and input the estimator obtained from the historic data set:

$$\boldsymbol{S}(m,k) = \sum_{t=m+1}^{m+k} H(\mathbf{X}_t, \widehat{\theta}_m),$$

Importantly, we allow the **monitoring function** H to differ from the estimating function G, where it is required that $\mathbb{E}H(\mathbf{X}, \theta_0) = 0$ with θ_0 defined by $\mathbb{E}G(\mathbf{X}, \theta_0) = 0$. Furthermore, the monitoring function H does not need to be a true estimating function (in the sense that $\mathbb{E}H(\mathbf{X}, \theta_0) = 0$ has a unique solution) but can map into a lower dimension $d' \leqslant d$. However, for $d' < d$ some alternatives are not asymptotically detectable while other are detected with a greater power. For a discussion of this effect for offline tests we refer to Kirch et al. [28] as well as Aston and Kirch [3]. The tests discussed in the literature so far, either use $H = G$ or the function that gives the estimated residuals in the respective model. The latter is usually (with the exception of tests considered by Ciuperca [12]) a projection of G, i.e. $H = G_1$ (with $G = (G_1, \ldots, G_d)$).

The heuristic behind this detector is the following: If no change occurs $\mathbb{E}H(\mathbf{X}_t, \widehat{\theta}_m) \approx \mathbb{E}H(\mathbf{X}_t, \theta_0) = 0$ for all t so that $\boldsymbol{S}(m,k)$ should be small. On the other hand if a change (in the best approximating parameters) occurs, then $\mathbb{E}H(\mathbf{X}_t, \widehat{\theta}_m) \approx \mathbb{E}H(\mathbf{X}_t, \theta_0) \neq 0$ for $t > k^*$, where k^* denotes the change point. Hence, under alternatives, $\boldsymbol{S}(m,k)$ will eventually have a trend away from 0. This different behavior can be statistically exploited to distinguish between the two alternatives. Since $\boldsymbol{S}(m,k)$ is possibly a vector, the corresponding monitoring scheme is based on a quadratic form, i.e. we reject as soon as

$$w^2(m,k)\boldsymbol{S}(m,k)^T\,\mathbf{A}\,\boldsymbol{S}(m,k) \geqslant c, \tag{2.2}$$

where $\boldsymbol{A}$ is a suitable symmetric positive (semi-)definite matrix, which can also be replaced by a consistent estimator. Here, c is a critical value, which can be derived from the asymptotics as discussed in Section 3, and $w(m,k)$ is a suitable weight function chosen in advance by the practitioner. As soon as (2.2) holds, we stop monitoring and reject the null hypothesis. Otherwise we continue monitoring.

We distinguish between **open-end procedures**, where we continue monitoring possibly to infinity, and **closed-end procedures** where we stop monitoring after a fixed number of observations $N(m)$ if the null hypothesis has not been rejected by then.

The statistical properties of this monitoring scheme can be described by the following stopping rule:

$$\tau(m) = \begin{cases} \inf\{1 \leqslant k < N(m) : w^2(m,k)\ \boldsymbol{S}(m,k)^T\boldsymbol{A}\boldsymbol{S}(m,k) \geqslant c\}, \\ \infty, \quad \text{if } w^2(m,k)\boldsymbol{S}(m,k)^T\,\mathbf{A}\,\boldsymbol{S}(m,k) < c, \text{ for all } 1 \leqslant k < N(m), \end{cases}$$

where $N(m) = \infty$ in case of an open-end procedure and $N(m) = Nm + 1$, $N > 0$, for the closed-end procedure. The stopping time $\tau(m)$ indicates the point in time at which we reject the null hypothesis and stop monitoring, where $\tau(m) = \infty$ indicates that the null hypothesis is never rejected.

Unlike in classical (nonsequential) statistics the sample size until a decision is reached is random and possibly infinite. Therefore, asymptotics with respect to the sample size tending to infinity are not suitable in this context. As a solution, Chu et al. [11] propose to use asymptotics with respect to the length m of the historic data set to grow to infinity. Since the historic data set is used for the parameter estimation of our model, this means in particular that this parameter estimation becomes better and better as $\widehat{\theta}_m \xrightarrow{P} \theta_0$.

As in standard statistical test procedures, our aim is to choose c such that we control the (asymptotic) α-error, i.e.

$$\lim_{m\to\infty} P_{H_0}\left(\tau(m) < \infty\right) = \alpha. \tag{2.3}$$

Theorem 3.1 shows how to choose the critical value c such that (2.3) holds, i.e. such that the procedure has asymptotic size α. Theorem 4.1 proves that the corresponding monitoring procedure detects a large class of alternatives with probability 1 asymptotically, i.e.

$$\lim_{m\to\infty} P_{H_1}\left(\tau(m) < \infty\right) = 1. \tag{2.4}$$

The choice of $w(m,k)$ determines the detection delay in dependence of the location of the change.

The detection power in such procedures is largely determined by the choice of estimating and monitoring function. The detection power is the better the more precise the estimators in (2.1) are as well as the better the monitoring function can distinguish between different parameter values. In this sense, using maximum likelihood scores will typically be preferable to using least squares scores. On the other hand, different properties such as robustness properties also carry over to the corresponding monitoring scheme leading to situations, where a more robust but less precise estimator can be preferable.

3 Null Asymptotics for sequential change-point tests

In this section, we derive the null asymptotics of the above sequential test statistics under certain regularity conditions on the estimating function and the observed process. In Section 6, we give examples where those regularity conditions are fulfilled.

The following assumptions impose certain regularity conditions on the weight function $w(m,k)$.

A. 1. a) The weight function is in the following class

$$w(m,k) = m^{-1/2}\tilde{w}\,(m,k)\,, \tag{3.1}$$

where $\tilde{w}(m,k) = \rho(k/m)$ for $k \geqslant a_m$ with $a_m/m \to 0$ and $\tilde{w}(m,k) = 0$ for $k < a_m$. The function ρ is continuous, and

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

b) For the open end procedure we additionally need

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

In particular, the conditions are fulfilled for

$$w(m,k) = m^{-1/2}\left(1+\frac{k}{m}\right)^{-1}\left(\frac{k}{m+k}\right)^{-\gamma} \tag{3.2}$$

with $0 \leqslant \gamma < 1/2$, which is the standard weight function proposed in the literature, because it leads to a nice asymptotic distribution for the open-end procedure (see Theorem 3.2 below).

Condition A.1 a) allows, e.g., for $w(m,k) = 0$ if $k \leqslant \log m$ without changing the asymptotic distribution. This may be useful as otherwise it can happen, that the false alarm rate right after monitoring starts is too high due to too few observations in $\boldsymbol{S}(m,k)$.

The choice of the weight function essentially determines the detection delay of the proposed procedure in dependence on the location of the change point. This is due to the fact that we stop if the partial sum process $\boldsymbol{S}(m,k)^T\boldsymbol{A}\boldsymbol{S}(m,k)$ crosses the boundary function $cw(m,k)^{-1}$ (for $w(m,k)\neq 0$) for the first time. Consequently, if we compare two procedures which only differ in the choice of boundary function (and the corresponding critical values), then for a given location the procedure with the lower boundary function (after having weighted it with the critical value) will have quicker detection delay for changes at that particular location. Two boundary function weighted with the critical value each for the same size (in the sense of (2.3)) will always cross each other at least once. Consequently, each will have a quicker detection delay for certain locations of change-points at the cost of a slower one for others.

The following set of assumptions guarantees that the estimator in the partial sum process can be replaced by the true (resp. best approximating) parameter θ_0 under the null hypothesis. It can be derived under suitable smoothness conditions on G and H using consistency properties of the estimator (confer Proposition 5.2 below).

The following approximation holds under H_0, where $N(m)$ is the possibly infinite observation horizon.

A. 2. For some θ_0 and $\boldsymbol{B}(\theta_0)$ depending on the distribution of $\mathbf{X}_1$,

$$\sup_{1\leqslant k<N(m)} w(m,k)\left\|\sum_{i=m+1}^{m+k} H(\mathbf{X}_i,\widehat{\theta}_m)\left(\sum_{j=m+1}^{m+k} H(\mathbf{X}_j,\theta_0)-\frac{k}{m}\boldsymbol{B}(\theta_0)\sum_{j=1}^{m} G(\mathbf{X}_j,\theta_0)\right)\right\| = o_P(1)$$

The dimension of the matrix $\boldsymbol{B}(\theta_0)$ guarantees that H and $\boldsymbol{B}(\theta)G$ are of the same dimension.

The additive term $\frac{1}{m}\boldsymbol{B}(\theta_0)\sum_{j=1}^m G(\mathbf{X}_j,\theta_0)$ accounts for the additional fluctuation of the first sum caused by the use of the estimator $\widehat{\theta}_m$ rather than the best approximating parameter θ_0. For sufficiently smooth estimating and monitoring functions, this condition can be derived by a Taylor expansion under weak moment conditions with

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

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 Section 5 below. In particular, whenever H is a linear combination of $G_1,\ldots,G_d$, then $\boldsymbol{B}(\theta_0)G(\mathbf{X},\theta_0)=H(\mathbf{X},\theta_0)$. This includes the standard situations that estimating and monitoring functions coincide or that the monitoring function is a projection onto one particular component of the estimating functions (such as estimated residuals for many estimating functions). Examples are discussed in detail in Section 6.

Many robust estimating functions are not differentiable, so that more involved considerations are necessary to show that the above condition is fulfilled (confer Example 6.2 below). On the other hand, for practical purposes those functions can be approximated to any degree of accuracy by sufficiently smooth functions.

The following assumptions control the fluctuations of the corresponding partial sum process:

A. 3. a) (i) For any $T > 0$ the partial sum process

$$\left\{ \frac{1}{\sqrt{m}} \sum_{t=1}^{\lfloor ms \rfloor} (H(\mathbf{X}_t, \theta_0), \boldsymbol{B}(\theta_0)\, G(\mathbf{X}_t, \theta_0)) : 1 \leqslant s \leqslant T \right\}$$

fulfills a functional central limit theorem towards a Wiener process $\{\mathbf{W}(s) : 1 \leqslant s \leqslant T\}$, $\mathbf{W}(s) = (\mathbf{W}_1(s), \mathbf{W}_2(s))$ with covariance matrix

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_1 & \mathbf{C} \\ \mathbf{C}^{\mathbf{T}} & \mathbf{\Sigma}_2 \end{pmatrix}. \tag{3.3}$$

(ii) The following Hájék-Rényi-type inequalities holds uniformly in m for all $0 < \alpha < 1/2$

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

(iii) For the open-end procedure the following Hájék-Rényi-type inequality is needed uniformly in m for any sequence $k_m > 0$

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

b) The partial sum process $\sum_{t=1}^{k} (H(\mathbf{X}_t, \theta_0), \boldsymbol{B}(\theta_0)\, G(\mathbf{X}_t, \theta_0))$ fulfills an invariance principle, i.e. (possibly after changing the probability space) there exists a Wiener process $\boldsymbol{W}$ with covariance matrix $\mathbf{\Sigma}$ as in (3.3) such that

$$\sum_{t=1}^{k} (H(\mathbf{X}_t, \theta_0), \boldsymbol{B}(\theta_0)\, G(\mathbf{X}_t, \theta_0)) - \boldsymbol{W}(k) = o_{\cancel{P}}(k^{1/2}) \text{ a.s. as } k \to \infty.$$

c) To obtain the extreme-value asymptotics in Theorem 3.1 c), we need that $\boldsymbol{B}(\theta)G = H$ and additionally that two independent Wiener processes $\{\mathbf{W}_1(\cdot)\}$ and $\{\mathbf{W}_2(\cdot)\}$ exist, each with covariance matrix Σ_1, such that for some $\nu > 2$ it holds

$$\sup_{k \geqslant 1} \frac{1}{k^{1/\nu}} \left(\sum_{t=m+1}^{m+k} H(\mathbf{X}_t, \theta_0) - \mathbf{W}_2(k) \right) = O_P(1),$$

$$\frac{1}{m^{1/\nu}} \left(\sum_{t=1}^{m} H(\mathbf{X}_t, \theta_0) - \mathbf{W}_1(m) \right) = O_P(1).$$

In the standard situation, where $\boldsymbol{B}(\theta_0)G(\mathbf{X}, \theta_0) = H(\mathbf{X}, \theta_0)$, these assumptions reduce to the corresponding assumptions on $H(\mathbf{X}, \theta_0)$. Typically, a) and even b) are obtained relatively easily, but c) is much harder in a dependent setting as it requires an argument leading to the asymptotic independence of the sums. Precisely, it involves some kind of cutting or big-block-small-block argument. Details for the proof in the special case of augmented GARCH time series can be found in Aue et al. [4].

Note that the invariance principle in (b) implies (a). To see this for (ii) and (iii) one needs to use the fact that by the stationarity of $\mathbf{X}_t$ it holds

$$\left\{\sum_{t=m+1}^{m+k} H(\mathbf{X}_t,\theta_0) : k \geqslant 1\right\} \overset{\mathcal{D}}{=} \left\{\sum_{t=1}^{k} H(\mathbf{X}_t,\theta_0) : k \geqslant 1\right\},$$

so that the invariance principle in addition to a standard Hájék-Rényi-inequality for i.i.d. normal data (applied to the increments of the Wiener process) yields the assertions. For the Darling-Erdős-type result the stronger rate of $o_P((\log\log m)^{-1/2})$ is needed.

Based on these regularity conditions, we can now prove the following null asymptotics:

Theorem 3.1. *Let Assumption A.2 and the null hypothesis hold.*

a) If Assumption A.1 (a) hold with $\tilde{w}(m,k) = \rho(k/m)$ for some bounded ρ as well as A.3 a) (i), then for any symmetric positive semi-definite $\mathbf{A}$, we get for the closed-end procedure

$$\lim_{m\to\infty} P\left(\sup_{1\leqslant k<Nm} w^2(m,k)\,\boldsymbol{S}(m,k)^T\,\mathbf{A}\,\boldsymbol{S}(m,k) \leqslant c\right)$$
$$= P\left(\sup_{0<t\leqslant N} \rho^2(t)(\boldsymbol{W}_1(t) - t\boldsymbol{W}_2(1))^T \boldsymbol{A}(\boldsymbol{W}_1(t) - t\boldsymbol{W}_2(1)) \leqslant c\right),$$

where $\{\mathbf{W}_1(t) : t \geqslant 0\}$ and $\{\mathbf{W}_2(t) : t \geqslant 0\}$ are independent Wiener processes with covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ respectively. For more general weight functions $\tilde{w}(m,k)$ as in A.1 the assertion remains true if additionally A.3 a) (ii) holds.

b) If Assumption A.1 (a) and (b) hold as well as A.3 a) (i) - (iii), then we get for the open-end procedure

$$\lim_{m\to\infty} P\left(\sup_{1\leqslant k<\infty} w^2(m,k)\,\boldsymbol{S}(m,k)^T\,\mathbf{A}\,\boldsymbol{S}(m,k) \leqslant c\right)$$
$$= P\left(\sup_{t>0} \rho^2(t)(\boldsymbol{W}_1(t) - t\boldsymbol{W}_2(1))^T \boldsymbol{A}(\boldsymbol{W}_1(t) - t\boldsymbol{W}_2(1)) \leqslant c\right),$$

where $\{\mathbf{W}_1(t) : t \geqslant 0\}$ and $\{\mathbf{W}_2(t) : t \geqslant 0\}$ are independent Wiener processes with covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ respectively. The supremum is well defined due to A.1.

c) If Assumptions A.2 with $B(\theta)G = H$ and the stronger rate $o_P((\log\log m)^{-1/2})$ and A.3 (c) hold for $w(m,k)$ as in (3.2) but with $\gamma = 1/2$, then the following Darling-Erdős theorem holds

$$\lim_{m\to\infty} P\left(a(\log m)\sup_{1\leqslant k<\infty} \frac{\sqrt{\boldsymbol{S}(m,k)^T\Sigma_1^{-1}\boldsymbol{S}(m,k)}}{\sqrt{m}\left(1+\frac{k}{m}\right)\left(\frac{k}{m+k}\right)^{1/2}} - b_{d'}(\log m) \leqslant t\right) = \exp(-e^{-t}),$$

where $a(x) = \sqrt{2\log x}, \qquad b_{d'}(x) = 2\log x + \frac{d'}{2}\log\log x - \log\Gamma(d'/2),$

$\Gamma(\cdot)$ *is the Gamma-function and* d' *the dimension of the monitoring function* H *i.e. of the vector* $\boldsymbol{S}(m,1)$.

The assertions remain true if $\mathbf{A}$ *is replaced by a consistent estimator for a) and b) and by an estimator fulfilling* $\|\widehat{\Sigma}_1^{-1/2} - \Sigma_1^{-1/2}\| = o_P((\log\log m)^{-1})$ *in c).*

In the standard situation $\boldsymbol{B}(\theta_0)G(\mathbf{X},\theta_0) = H(\mathbf{X},\theta_0)$ it holds $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ and a canonical choice for the matrix $\mathbf{A}$ is given by $\boldsymbol{\Sigma}_1^{-1}$. For this choice the limit distribution is pivotal in the sense that it does not depend on any unknowns.

The following theorem shows that for particular weight functions, the limit in the open-end procedure can be simplified. Part a) is well known and is the main reason why the weight functions in (3.2) are so popular for the open-end procedure.

Theorem 3.2. *a) If* $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, *then for any* $0 \leqslant \gamma < 1/2$

$$\sup_{t>0} \frac{(\boldsymbol{W}_1(t) - t\boldsymbol{W}_2(1))^T \boldsymbol{A}(\boldsymbol{W}_1(t) - t\boldsymbol{W}_2(1))}{(1+t)^2 \left(\frac{t}{1+t}\right)^{2\gamma}} \overset{\mathcal{D}}{=} \sup_{0<t<1} \frac{\boldsymbol{W}(t)^T \boldsymbol{A}\boldsymbol{W}(t)}{t^{2\gamma}},$$

where $\{\boldsymbol{W}(\cdot)\}$ *is a Wiener process with covariance matrix* $\boldsymbol{\Sigma}_1$.

b) If $\boldsymbol{\Sigma}_1 = \sigma_1^2 \neq \sigma_2^2 = \boldsymbol{\Sigma}_2$, *then*

$$\sup_{t>0} (\sigma_2^2)^{1-2\gamma} \frac{(W_1(t) - tW_2(1))^2}{(\sigma_1^2 + \sigma_2^2 t)^2 \left(\frac{t}{\sigma_1^2+\sigma_2^2 t}\right)^{2\gamma}} \overset{\mathcal{D}}{=} \sup_{0\leqslant t\leqslant 1} \frac{W^2(t)}{t^{2\gamma}},$$

where $\{W(\cdot)\}$ *is a univariate standard Wiener process.*

4 Consistency under alternatives

In this section, we state some regularity conditions under which the procedure eventually stops with probability one if a change does occur, which means that it has asymptotic power one. In particular, we concentrate on fixed change alternatives, where a fixed parameter θ_1 not depending on m is used after the change. The effect of the size of change on power or detection delay cannot be quantified by such changes but asymptotics for local changes, where $\theta_1 = \theta_{1,m} \to \theta_0$ need to be considered. In the linear case, this is usually not difficult, however, in non-linear situations this requires even more technical assumptions and therefore will not be considered here. For a detailed analysis in a non-linear retrospective situation, see Kirch and Tadjuidje Kamgaing [32].

As discussed below Condition A.1 the detection delay time for a given set of estimating/monitoring functions is determined mainly by the weight function that is used in combination to the location of the change, where it typically takes longer for later changes to be detected. For early local change scenarios and weight functions as in (3.2) Aue and Horváth [6] and Aue et al. [8] derive the asymptotic distribution of the stopping time in a mean change model respectively a linear regression model.

A. 4. a) The time series before the change fulfills the assumptions under the null hypothesis.

b) The change-point is of the form $k^* = \lfloor m\vartheta \rfloor$ for some $0 < \vartheta < N$ (where $N = \infty$ in case of the open-end procedure). Furthermore, there exist a vector $\mathbf{E}_H$ and a ball $U(x_0)$ around x_0 with $x_0 > \vartheta$ and $\inf_{x \in U(x_0)} \rho(x) > 0$ such that

$$\frac{1}{m} \left\| \sum_{j=m+k^*+1}^{\lfloor x_0 m \rfloor} \left(H(\mathbf{X}_j, \widehat{\theta}_m) - \mathbf{E}_H \right) \right\| = o_P(1). \tag{4.1}$$

c) In the open-end procedure we can allow for an arbitrarily late change k^* if $\liminf_{x \to \infty} x\rho(x) > 0$ as well as if for $l \to \infty$ there exists a vector $\mathbf{E}_H$ such that

$$\frac{1}{l} \left\| \sum_{j=m+k^*+1}^{m+k^*+l} \left(H(\mathbf{X}_j, \widehat{\theta}_m) - \mathbf{E}_H \right) \right\| = o_P(1). \tag{4.2}$$

d) It holds $\mathbf{A}^{1/2}\mathbf{E}_H \neq 0$, where $\mathbf{A}$ is the matrix used (or estimated) in the monitoring procedure.

Condition $\mathbf{A}^{1/2}\mathbf{E}_H \neq 0$ is the key to which alternatives are detectable. If the time series $\{X^*(t)\}$ after the change is stationary and ergodic (or sufficiently close to it), then typically $\boldsymbol{E}_H = \mathbb{E}H(\mathbf{X}_1^*, \theta_0)$. If the monitoring function is an estimating function in the sense that $\mathbb{E}H(\mathbf{X}_1, \theta_0) = 0$ has a unique solution, then $\mathbb{E}H(\mathbf{X}_1^*, \theta_0) \neq 0$ whenever the time series before and after the change have different best approximating parameters (in the sense of G and H respectively). However, if $d' < d$, then not all changes in the best approximating parameters can be detected, but some alternatives can be detected with better power (confer Kirch et al. [28] resp. Aston and Kirch [3] for a detailed discussion of this effect in the offline case).

Assumptions (4.1) and (4.2) can be obtained for sufficiently smooth estimating functions under weak moment conditions, see Section 5.2 below.

Theorem 4.1. *Under Assumptions A.4 a), b) and d), the closed-end procedure has asymptotic power one, hence will eventually stop. The open-end procedure has asymptotic power one under A.4 a),d) and either b) or c). This remains true if a consistent estimator for* $\mathbf{A}$ *is used.*

In offline procedures the estimator for $\mathbf{A}$ is typically contaminated under alternatives exhibiting a different limit behavior than under the null hypothesis. In the sequential setting, however, such an estimator is based on the historic data set only, so that it will have the same behavior under both the null hypothesis as well as alternatives.

5 Regularity conditions for sufficiently smooth functions

In this section we give some moment conditions for which we can prove the main regularity assumptions of the previous section as long as the estimating and monitoring function are sufficiently smooth. While this is a general approach suitable in many situations, some robust estimating functions of interest (such as L_1-minimizer) are not smooth. For important special cases of those functions the regularity conditions of the previous sections have been proven using different methods (confer Section 6.2 below). On the other hand, such estimating functions can often be approximated to any degree of accuracy by estimating functions fulfilling the smoothness conditions stated here.

5.1 Conditions under the null hypothesis

Under the below assumptions we can derive Assumption A.2.

B. 1. Let $\{\mathbf{X}_t\}$ be stationary and ergodic under the null hypothesis.

B. 2. a) $\mathbb{E}\sup_{\theta\in\Theta}\|G(\mathbf{X}_1,\theta)\| < \infty$.

b) θ_0 is the unique zero of $\mathbb{E}G(\mathbf{X}_1,\theta)$ in the strict sense, i.e. for every $\epsilon > 0$ there exists a $\delta > 0$ such that $\|\mathbb{E}G(\mathbf{X}_1,\theta)\| > \delta$ whenever $\|\theta-\theta_0\| > \epsilon$.

c) G is continuously differentiable with respect to θ in a convex environment U_{θ_0} of θ_0 (for θ_0 as in b)) such that $\mathbb{E}\nabla G(\mathbf{X}_1,\theta_0)$ is positive definite and

$$\mathbb{E}\sup_{\theta\in U_{\theta_0}}\|\nabla G(\boldsymbol{X}_1,\theta)\| < \infty.$$

d) $\sum_{j=1}^{m} G(\mathbf{X}_t,\theta_0) = O_P(\sqrt{m})$.

The first assertions are regularity conditions, while the last follows from a central limit theorem for $G(\mathbf{X}_t,\theta_0)$ under weak moment conditions in addition to weak dependence assumptions.

Proposition 5.1. *a) Under Assumptions B.1 and B.2 a) and b) it holds* $\widehat{\theta}_m \xrightarrow{P} \theta_0$.

b) Under Assumptions B.1 and B.2 it holds $\widehat{\theta}_m = \theta_0 + O_P(m^{-1/2})$.

In order to derive Assumption A.2 we need the assertion of Proposition 5.1 b) in addition to some additional regularity conditions on both H and G.

B. 3. Denote for $F = (F_1,\ldots,F_d)$ the gradient matrix $\nabla F = (\nabla F_1,\ldots,\nabla F_d)^T$, where ∇ is the gradient (with respect to θ). Then we assume

$$\mathbb{E}\|\nabla H(\mathbf{X}_1,\theta_0)\| < \infty.$$

Furthermore, for $j = 1,\ldots,d'$, it holds for some convex environment U_{θ_0} of θ_0

$$\mathbb{E}\sup_{\xi\in U_{\theta_0}}\|\nabla^2 H_j(\mathbf{X}_1,\xi)\|_\infty < \infty, \qquad \mathbb{E}\sup_{\xi\in U_{\theta_0}}\|\nabla^2 (B(\theta_0)\,G)_j(\mathbf{X}_1,\xi)\|_\infty < \infty.$$

Proposition 5.2. *Under Assumptions B.1, B.2 and B.3 with*

$$\boldsymbol{B}(\theta_0) = \mathbb{E}\nabla H(\mathbf{X}_0,\theta_0)\ \left(\mathbb{E}\nabla G(\mathbf{X}_0,\theta_0)\right)^{-1}, \tag{5.1}$$

Assumption A.2 follows for weight functions fulfilling A. 1 a) for the closed-end in addition to b) for the open-end procedure.

5.2 Conditions under alternatives

In this section we give some moment conditions for which we derive A.4 under alternatives. To this end, let us denote by $m + k^*$ the time when a change occurs.

B. 4. It holds $\boldsymbol{X}_t = \boldsymbol{X}_t^*$ for $t > m + k^*$ for a stationary and ergodic time series $\{\boldsymbol{X}_t^*\}$ such that for some convex environment U_{θ_0} of θ_0 it holds

$$\mathbb{E}\|\nabla H(\boldsymbol{X}_1^*, \theta_0)\| < \infty, \qquad \mathbb{E} \sup_{\xi \in U_{\theta_0}} \|\nabla^2 H_j(\boldsymbol{X}_1^*, \xi)\|_\infty < \infty.$$

Obviously, the moment conditions required for the time series after the change are much weaker than the ones required for the time series before the change as given in B.2 and B.3.

Proposition 5.3. *Under Assumptions B.4 it holds* (4.1) *and* (4.2) *with* $\boldsymbol{E}_H = \mathbb{E}H(\boldsymbol{X}_1^*, \theta_0)$.

In particular in autoregressive models the stationarity assumption in B.4 is often too strong as starting values from the time series before the change are to be expected. In this case, the following assumptions can help:

B. 5. a) The time series after the change point $m+k^*$ can be written as $\boldsymbol{X}_t = \boldsymbol{X}_t^* + \boldsymbol{R}_t$, where $\boldsymbol{X}_t^*$ fulfills B.3 and as $l \to \infty$

$$\frac{1}{l} \sum_{t=m+k^*+1}^{m+k^*+l} \|\boldsymbol{R}_t\|^2 = o_P(1).$$

b) For $j > m + k^*$ it holds $\|H(\mathbf{X}_j, \widehat{\theta}_m) - H(\mathbf{X}_j^*, \widehat{\theta}_m)\| \leqslant \|\boldsymbol{R}_j F(\boldsymbol{X}_j^*)\| + C\|\boldsymbol{R}_j\|^2$ for some measurable function F such that $\mathbb{E}\|F(\boldsymbol{X}_j^*)\|^2 < \infty$.

Assumption a) allows for starting values from a different distribution as long as the difference to the time series with starting values from the stationary distribution is small enough. A similar idea has also been used in Horváth et al. [24]. For linear autoregressive models this is naturally fulfilled and can easily be checked, for non-linear autoregressive settings, some work needs to be done.

Proposition 5.4. *Under Assumption B.5 we get* (4.1) *and* (4.2) *with* $\boldsymbol{E}_H = \mathbb{E}H(\boldsymbol{X}_1^*, \theta_0)$.

The following proposition gives some conditions under which Assumption B.5 holds. Markov chains that are geometric ergodic are also mixing in the below ergodic theoretic sense (confer the definition in Meyn and Tweedie [39] in addition to condition (iv) in Theorem 21.12 in Lindvall [38]).

Proposition 5.5. *For a mixing (in the ergodic theoretic sense) Markov chain* $\{\boldsymbol{X}_t\}$ *which starts in* $\boldsymbol{X}_0 = \mathbf{x}_0$ *(not necessarily from the stationary distribution) there exists a stationary process* $\{\boldsymbol{X}_t^*\}$ *and a random process* $\{\boldsymbol{R}_t\}$*, such that* $\boldsymbol{X}_t = \boldsymbol{X}_t^* + \boldsymbol{R}_t$ *and as* $l \to \infty$

$$\frac{1}{l} \sum_{t=1}^{l} \|\boldsymbol{R}_t\|^2 = o_P(1).$$

6 Examples and Simulation Studies

In the following subsections we present a survey of existing results and provide new examples that have not yet been discussed in the literature. Because of the generality of the weight functions considered in this paper, we extend existing results that have often only been obtained for weight functions as in (3.2). To illustrate the small sample behavior of the proposed procedures, some of the examples are accompanied by simulations and data examples. The empirical results of the simulations are always based on 1000 repetitions. The critical values are chosen according to the limit distributions derived in the previous sections, where we used simulated critical values based on 10 000 repetitions except for the extreme-value asymptotics.

6.1 Linear regression

Consider the classical linear regression model

$$X_t = \mathbb{Z}_t^T \boldsymbol{\beta}_0 + \varepsilon_t, \tag{6.1}$$

where $\boldsymbol{\beta}$ is the unknown regression parameter, and $\{\mathbb{Z}_t\}$ are random regressors with $\mathbb{Z}_t = (1, Z_{t,2}, \dots, Z_{t,p})^T$ independent of $\{\varepsilon_t\}$ and fulfilling for some positive definite matrix $\boldsymbol{C}$ and $\tau > 0$ that

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{Z}_i \mathbb{Z}_i^T - \boldsymbol{C} = O(n^{-\tau}) \qquad a.s. \tag{6.2}$$

The papers of Chu et al. [11] and Horváth et al. [23] on this example with independent errors $\{\varepsilon_i\}$ has triggered the development of the above methodology. They propose to use the ordinary least squares estimator as estimating function, i.e.

$$G((X_t, \mathbb{Z}_t^T), \boldsymbol{\beta}) = \mathbb{Z}_t(X_t - \boldsymbol{\beta}^T \mathbb{Z}_t).$$

The monitoring function is given by the estimated residuals, i.e.

$$H((X_t, \mathbb{Z}_t^T), \boldsymbol{\beta}) = X_t - \boldsymbol{\beta}^T \mathbb{Z}_t.$$

Since by assumption $Z_{t,1} = 1$, the monitoring function H is the first line of G, hence we get $\boldsymbol{B}(\boldsymbol{\beta}_0)G = H$. Because this is a real function the monitoring rule in (2.2) can equivalently be expressed by

$$|\mathbf{S}(m,k)|/\sigma \geqslant \tilde{c}/w(m,k),$$

where $\tilde{c}$ is the quantile based on the asymptotic distribution of this CUSUM detector. This is the form that was analyzed in Chu et al. [11] and Horváth et al. [23].

Horváth et al. [23] prove Assumption A.2 with $\theta_0 = \boldsymbol{\beta}_0$ in their Lemma 5.2 for weight functions as in (3.2) with $0 \leqslant \gamma < \min(\tau, 1/2)$, but their proof remains true for weight functions as in A.1 as long as $\gamma < \tau$ in A.1 a). Since $H((X_t, \mathbb{Z}_t^T), \boldsymbol{\beta}_0) = \varepsilon_t$ under H_0, Condition A.3 simplifies to the corresponding assumptions on the error terms. In the case of i.i.d. errors A.3 c) follows from the invariance principles by Komlós et al. [34, 35]. Extensions to the non-i.i.d. case have been proposed by Aue et al. [7], for certain martingale difference sequences including augmented GARCH processes, as well as by Schmitz and Steinebach [43], for certain weak dependent processes. Horváth et al. [25] prove the corresponding Darling-Erdősz-result as given in Theorem 3.1 c) above.

Due to the fact that the monitoring function H is not a full estimating function, restrictions apply which alternatives are detectable. In fact, the proof of Theorem 2.2 in Horváth et al. [23] shows (4.2) with $\boldsymbol{E}_H = \boldsymbol{c}_1^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)$ if $X_t = \mathbb{Z}_t^T\boldsymbol{\beta}_1 + \varepsilon_t$ after the change, where $\boldsymbol{c}_1$ is the first column of the matrix $\boldsymbol{C}$ as in (6.2). Consequently, only changes are detectable for which $\boldsymbol{E}_H \neq 0$ which means that the change goes along with a mean change.

Because of the power restriction above, Hušková and Koubková [26] proposed to use the monitoring function

$$H((X_t, \mathbb{Z}_t^T), \boldsymbol{\beta}) = \mathbb{Z}_t(X_t - \boldsymbol{\beta}^T\mathbb{Z}_t) \tag{6.3}$$

for proving the above assumptions. For this choice, all alternatives have asymptotic power one because $\boldsymbol{E}_H = \mathbf{C}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) \neq 0$ due to the positive definiteness of $\mathbf{C}$.

6.2 Mean changes

For $\mathbb{Z}_t = 1$ model (6.1) reduces to a mean change model, while the estimation and monitoring is related to the sample mean.

Because the mean is not robust, extensions to M-estimators have been considered in the dissertations of Koubková [36] as well as Chochola [10]. Koubkova [36] (Lemmas 5.1 and 5.3) shows in particular that A.2 holds for the L_1-procedure where $G(X_i, \mu) = H(X_i, \mu) = \operatorname{sgn}(X_i - \mu)$ (leading to the median as estimator). Additionally, she considers more general M-estimating equations, where she proves A.2 in (6.14). Assumptions A.3 can be obtained by standard methods on the i.i.d. errors $\operatorname{sgn}(X_i - \mu_0)$ respectively $\psi(X_i - \mu_0)$. Koubková [36] also considers the linear regression situation of the previous subsection. Chochola [10] considers M-estimating equations for monitoring and proves A.2 in his Lemma 2.6. Additionally, he gives extensions to the multivariate location model.

Since both Koubková [36] and Chochola [10] consider general M-estimators, which are not necessarily differentiable (or even continuous), the methodology provided by Proposition 5.2 cannot be applied and a different approach is necessary to obtain A.2.

Because our theory allows for the estimating and monitoring functions to differ, we can combine a more precise estimator based on the historic data with a more robust monitoring function. Such a procedure can be important in practice if, for example, the historic data set has no outliers but the newly arriving observations are likely to exhibit some. Since typically the historic data set is relatively small in comparison to the possible length of the observation horizon, getting a more precise estimator from the historic data set may be crucial. The below choice of monitoring functions fulfills the regularity assumptions of Proposition 5.2. For other less smooth choices as e.g. the sign-function corresponding to the median, some additional work is required in order to derive A.2.

Simulation study

For illustrational purposes we will simulate data according to $X_t = \mu + \varepsilon_t$ for i.i.d. errors ε_t (possibly contaminated by outliers after monitoring starts). We will initially use the (non-robust) sample mean based on the historic data set, i.e. $G(X,\mu) = X - \mu$, but then use the following more robust monitoring function (which estimates the mean for symmetric data):

$$H(X,\mu) = \tanh(X-\mu), \qquad \tanh(u) = \frac{e^u - e^{-u}}{e^u + e^{-u}}.$$

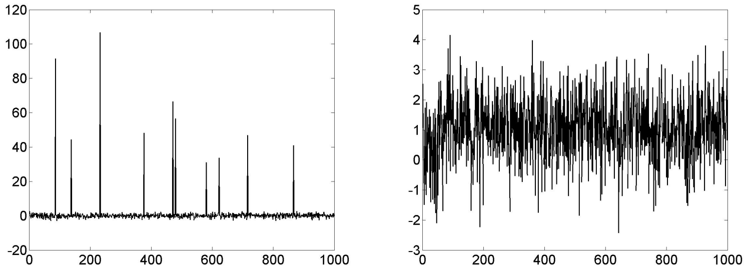
By $\frac{\partial \tanh(u)}{\partial u} = 1 - \tanh^2(u)$ and Proposition 5.2 it follows $\mathbf{B}(\mu_0) = \mathbb{E}\tanh^2(X-\mu_0) - 1$, so that we get by Theorem 3.1 b) and Theorem 3.2 b)

$$\sup_{1\leq k<\infty} \frac{1}{\sqrt{m}} \frac{\sigma_2}{\sigma_1^2 + \sigma_2^2 \frac{k}{m}} \left| \sum_{j=m+1}^{m+k} \tanh(X_j - \bar{X}_m) \right| \xrightarrow{\mathcal{D}} \sup_{0\leqslant t\leqslant 1} |W(t)|,$$
$$\text{where } \sigma_1^2 = \operatorname{var}\tanh(X_1-\mu_0), \qquad \sigma_2^2 = (\mathbf{B}(\mu_0))^2 \operatorname{var}(X_1-\mu_0).$$

In the simulations, we replace σ_j by consistent estimators given by

$$\widehat{\sigma}_1^2 = \frac{1}{m}\sum_{j=1}^m \tanh^2(X_j - \bar{X}_m) - \left(\frac{1}{m}\sum_{j=1}^m \tanh(X_j - \bar{X}_m)\right)^2,$$
$$\widehat{\sigma}_2^2 = \left(\frac{1}{m}\sum_{j=1}^m \tanh^2(X_j - \bar{X}_m) - 1\right)^2 \left(\frac{1}{m}\sum_{j=1}^m (X_j - \bar{X}_m)^2\right).$$

We now apply the monitoring scheme to the null data set with standard normal errors (H_0), standard normal errors with no change in the mean but outliers (H_c), where 1% of the random variables have been randomly replaced by $\Gamma(5,10)$ observations, and two time series with independent standard normal errors and a mean change of magnitude ($d = 1$, $d = -0.5$ at $k^* = \frac{m}{2}$). In Figure 6.1 a sample path for each of the latter three time series is given.



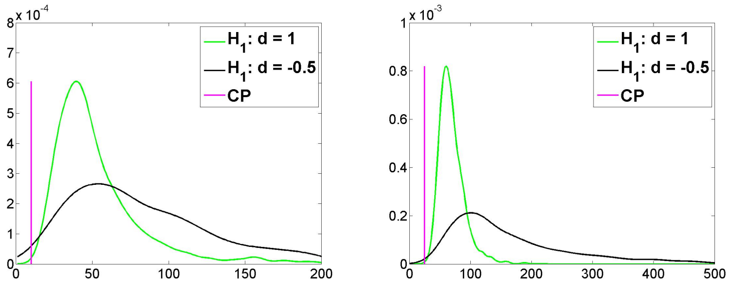
(a) Null data with outliers (H_c) (b) $H_1 : k^* = \frac{m}{2}, d = 1$ (c) $H_1 : k^* = \frac{m}{2}, d = -0.5$

Figure 6.1: Mean change: Sample of monitoring data ($m = 100$)

In the simulations, we stop monitoring after $10\,m$ observations. The empirical size and power for the above setting are reported in Table 6.1 and Figure 6.2 shows a scaled density estimator for the run length, i.e. the time until the procedure stops, which is scaled to integrate to the empirical level. The vertical line indicates the change point.

	H_0			H_c			$H_1 : d = 1$			$H_1 : d = -0.5$		
m	20	50	100	20	50	100	20	50	100	20	50	100
	0.027	0.034	0.040	0.031	0.039	0.037	0.996	1.00	1	0.406	0.823	0.991

Table 6.1: Mean change: Empirical size and power for a nominal 5% level (H_0: no change (standard normal), H_c: no change but outliers, H_1: mean changes of magnitude d at $k^* = \frac{m}{2}$)



(a) $m = 20$ (b) $m = 50$ (c) $m = 100$

Figure 6.2: Mean Change: Scaled density estimate of the run length for mean changes of magnitude $d = 1$ resp. $d = -0.5$. The vertical line indicates where the change point occurs.

The procedure is conservative with respect to the size for both situations with and without outliers but detects changes in the mean quite well, where for smaller changes a longer historic data set is needed in order to get a good detection rate.

6.3 Non-linear models

Several applications to non-linear time series have already been discussed in the literature. Berkes et al. [9] use the log likelihood score function as estimating as well as monitoring function to detect changes in GARCH parameters. Their Lemma 6.4 proves Assumption 2, the proof of Lemma 6.6 shows A.3 (i) – (iii).

Ciuperca [12] considers a nonlinear regression model $X_i = f(\mathbb{Z}_i, \boldsymbol{\beta}_0) + \varepsilon_i$ with known regression function f and i.i.d. errors $\{\varepsilon_i\}$. For her initial estimation of the parameter $\boldsymbol{\beta}_0$ she uses the ordinary least squares estimator with estimating function

$$G((X_t, \mathbb{Z}_t), \boldsymbol{\beta}) = \nabla f(\mathbb{Z}_t, \boldsymbol{\beta})(X_t - f(\mathbb{Z}_t, \boldsymbol{\beta})).$$

Her monitoring function is based on estimated residuals, i.e.

$$H((X_t, \mathbb{Z}_t), \boldsymbol{\beta}) = X_t - f(\mathbb{Z}_t, \boldsymbol{\beta}).$$

Since this is not an estimating function, she cannot expect to detect all changes even in the correctly specified model, where the parameter $\boldsymbol{\beta}_0$ changes to $\boldsymbol{\beta}_1$ after the change. In fact the proof of her Theorem 3.2 shows that (4.2) holds with $\boldsymbol{E}_H = \mathbb{E}f(\mathbb{Z}_1, \boldsymbol{\beta}_0) - \mathbb{E}f(\mathbb{Z}_1, \boldsymbol{\beta}_1)$. Thus only changes can be detected that go along with a mean change. She uses an open-end as well as a closed-end procedure with the standard weight function in (3.2). Unlike for linear regression the monitoring function H is no longer , necessarily, a linear combination of the components of G such that $\boldsymbol{B}(\theta_0)G \neq H$. In this situation the weight function (3.2) does no longer lead to a pivotal limit with all the problems this entails. On the other hand a weight function as in Theorem 3.2 b) leads to a pivotal limit. Since the regularity conditions of Ciuperca [12] are similar to the ones given in Section 5, $\boldsymbol{B}(\theta_0)$ is granted by (5.1). Condition A.2 is proven in her Lemma A.1, Conditions A.3 then follows by the standard invariance principle of Komlos et al. [34, 35].

Neural network based detectors for non-linear autoregressive time series

Kirch and Tadjuidje Kamgaing [29] use a similar monitoring scheme for non-linear nonparametric autoregressive time series. Similarly to the approach by Kirch and Tadjuidje Kamgaing [30] in the offline setting, a detector is used that detects changes in the best approximating parameters of an autoregressive time series where the regression function is given by a neural network. Since neural network functions can approximate a large class of functions to any degree of accuracy (confer e.g. White [49] or Franke et al. [18] and some of the references therein), this is similar to the idea behind sieve estimators in nonparametric statistics. Because it cannot be expected that the time series follows that precise model, misspecification is then taken into account.

To elaborate, assume that we observe X_t with

$$X_t = g(X_{t-1}, \dots, X_{t-p}) + \varepsilon_t, \tag{6.4}$$

which is stationary and ergodic with existing fourth moments and also strong mixing with exponential rate.

In order to construct a detector for a change in the autoregression function g, we use a parametric approximation based on a one layer feedforward neural network with n_H hidden neurons

$$f(\mathbf{x}, \theta) = \nu_0 + \sum_{h=1}^{n_H} \nu_h \psi(< \boldsymbol{\alpha}_h, \mathbf{x} > + \beta_h), \tag{6.5}$$

where $\theta = (\nu_0, \dots, \nu_{n_H}, \boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_{n_H}, \beta_1, \dots, \beta_{n_H}) \in \Theta$ and we assume Θ to be convex and compact, $\boldsymbol{\alpha}_j = (\alpha_{j1}, \dots, \alpha_{jp})$ and $<,>$ is the classical scalar product on $\mathbb{R}^p$. Furthermore, we assume that ψ is twice continuously differentiable with bounded first and second derivatives and belongs to the class of sigmoid activation functions that satisfy

$$\lim_{x\to-\infty} \psi(x) = 0, \qquad \lim_{x\to\infty} \psi(x) = 1, \qquad \psi(x) + \psi(-x) = 1. \tag{6.6}$$

A popular example is the logistic function $\psi(x) = (1 + e^{-x})^{-1}$.

We use an ordinary least-squares approach to estimate the best approximating parameters, i.e. we use

$$G((X_t, \mathbb{X}_{t-1}), \theta) = (X_t - f(\mathbb{X}_{t-1}, \theta)) \, \nabla f(\mathbb{X}_{t-1}, \theta), \ \mathbb{X}_{t-1} = (X_{t-1}, \dots, X_{t-p})^T, \tag{6.7}$$

as estimating function which has dimension $d = n_H(p + 2) + 1$. Kirch and Tadjuidje Kamgaing [29] consider monitoring schemes based on the estimated residuals, i.e. they use as monitoring function

$$H((X_t, \mathbb{X}_{t-1}), \theta) = X_t - f(\mathbb{X}_{t-1}, \theta) \tag{6.8}$$

with $\boldsymbol{B}(\theta_0)G = H$.

Alternatively, we can use the full estimating function in the monitoring procedure, i.e. $\tilde{H} = G$ (and $\boldsymbol{B}(\theta_0) = \mathrm{Id}$).

Regularity conditions

In the above situation Condition B.1 is fulfilled. Condition A.3 b) follows for (G, H) as well as $(G, \tilde{H})$ from the invariance principle of Kuelbs and Philipp [37] for mixing random variables, while c) can be obtained from the mixing assumption in addition to a big block small block argument. The mixing assumption is only used to keep the arguments simple, however the use of other weak dependency concepts implying Conditions A.3 is also possible. For many nonlinear autoregressive models as in (6.8) one can make use of standard Markov chain stability theory (see e.g. Meyn and Tweedie [39] or Tong [47]) to prove that the process X_t is geometric ergodic. The latter property implies the existence of a unique (asymptotic) stationary solution for X_t which satisfies the absolute regularity property as well, i.e. β-mixing with exponential rate. The proof of the geometric ergodicity is based on ϕ-irreducibility, aperiodicity and the drift condition for Markov chains, that need to be guaranteed (see Chapter 15 of Meyn and Tweedie [39]). Such application can be found in Stockis et al. [46], in a broader framework, where they use autoregressive time series based on neural network functions as building blocks in a regime-switching model, so called CHARME-models. We are interested in monitoring for changes in the autoregression function g.

Since f is twice continuously differentiable by assumption, G is continuously differentiable with respect to θ. By the boundedness of ψ and its first derivative and the compactness of Θ, Condition B.2 a) follows if $\{X_t\}$ is square-integrable, while the existence of the moment in c) follows if at least the third moments exist. The latter also implies Assumption B.2 d) by the mixing property. It remains to assume B.2 b), which is essentially an identifiability condition, saying that the best approximating parameter θ_0 given by $\mathbb{E}G((X_t, \mathbb{X}_{t-1}, \theta_0) = 0$ is identifiable unique. The positive definiteness condition in c) is another regularity condition which is standard in the literature, see for example Hall [21].

Consequently, by Proposition 5.1 we get $\sqrt{m}$-consistency of the least squares estimator to the best approximating parameter.

B.3 is fulfilled for H due to the existence of second moments and the boundedness of the first and second derivatives of the activation function ψ in addition to the compactness assumption of Θ. For $\widetilde{H}$ B.3 follows if the activation function ψ is additionally three times continuously differentiable with bounded third derivative due to the existence of fourth moments.

By Proposition 5.2 we get A.2.

Monitoring statistics

Under the above assumptions, Theorem 3.1 is applicable based on both monitoring functions H as well as $\tilde{H}$, so that a large class of detectors is available. If one chooses $\boldsymbol{A}$ as the inverse of the long-run covariance matrix of $H((X_t, \mathbb{X}_{t-1}), \theta_0)$ respectively $\tilde{H}((X_t, \mathbb{X}_{t-1}), \theta_0)$ one gets pivotal limits where the components of the Wiener processes are independent. In the correctly specified causal model with independent errors, this long-run variance reduces to the error variance σ^2 in case of H and to

$$\sigma^2\, \mathbb{E}\nabla f((X_t, \mathbb{X}_{t-1}), \theta)\, \nabla f((X_t, \mathbb{X}_{t-1}), \theta)^T$$

in case of $\tilde{H}$. We propose to use

$$\widehat{\mathbf{A}}_H = \widehat{\sigma}^{-2}, \qquad \text{with} \quad \widehat{\sigma}^2 = \frac{1}{m - n_H(p+2) + 1} \sum_{j=p+1}^{m} (X_j - f((X_j, \mathbb{X}_{j-1}), \widehat{\theta}_m))^2,$$

respectively

$$\widehat{\mathbf{A}}_{\tilde{H}} = \widehat{\sigma}^{-2} \left(\frac{1}{m - n_H(p+2) + 1} \sum_{j=p+1}^{m} \nabla f((X_j, \mathbb{X}_{j-1}), \widehat{\theta}_m)\, \nabla f((X_j, \mathbb{X}_{j-1}), \widehat{\theta}_m)^T \right)^{-1}.$$

As suggested by the simulations in Kirch and Tadjuidje Kamgaing [30] for the offline case, this choice often leads to better results than an estimator for the long-run variance.

	$H_0 : \theta_0 = (1, 0.3)$			$H_1 : \theta_1 = (3, 0.75)$			$H_1 : \theta_1 = (1, 0.75)$			$H_1 : \theta_1 = (3, 0.3)$		
m	100	200	300	100	200	300	100	200	300	100	200	300
	0.065	0.045	0.036	0.979	0.994	0.996	0.684	0.950	0.996	0.991	0.993	0.999

Table 6.2: Empirical size and power (nominal 5% level) for the misspecified neural network monitoring with a true underlying AR(1)-time series and the extreme value type statistic, where θ_0 indicates the parameter under the null hypothesis resp. before a change occurs and θ_1 indicate the parameters after the change has occurred

This trades a large estimation error for the long-run variance with a small model error because the procedure is only applied if the fit based on a neural network is reasonably good for the historic data set.

Behavior under alternatives

In order to understand the behavior of the two statistics under alternatives better, note that by the mean value theorem, the boundedness of the first derivative of ψ and the compactness of Θ, it holds

$$\sup_{\theta \in \Theta} \|H(\mathbf{x}, \theta) - H(\mathbf{y}, \theta)\| \leqslant D\|\mathbf{x} - \mathbf{y}\|$$

for a suitable constant D.

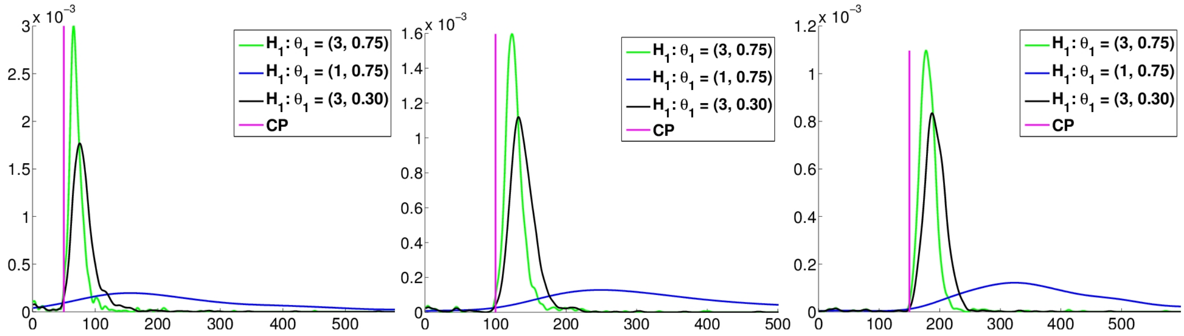
Furthermore, since the first derivative of ψ is bounded and Θ is compact, we get

$$\sup_{\theta \in \Theta} \|\nabla f(\mathbf{x}, \theta) - \nabla f(\mathbf{y}, \theta)\| \leqslant D(\|\mathbf{y}\| + \|\mathbf{x} - \mathbf{y}\|).$$

This implies B.5 b) for both H (with $F = D$) as well as $\tilde{H}$ (with $F = D\,\mathrm{id}$) if $\{X_t^*\}$ has second moments. In this case $\boldsymbol{E}_H = \mathbb{E}(X_t^*) - \mathbb{E}f(\mathbb{X}_t^*, \theta_0)$, so that essentially mean changes will be detected. On the other hand $\boldsymbol{E}_{\tilde{H}} = \mathbb{E}G(X_t^*, \theta_0)$ which is different from 0 as soon as the best approximating parameters exist and differ for both time series. Consequently, only the procedure based on $\tilde{H}$ has asymptotic power one for all changes leading to different best approximating parameters in the neural network approximation.

Simulation study

We apply the detector based on the estimating and monitoring function as in (6.7) resp. (6.8) with $\widehat{\mathbf{A}}_H$ (and a neural network as in (6.5) with 2 hidden neurons) to a linear autoregressive process $X_t = \omega + \alpha X_{t-1} + \varepsilon_t$, $\{\varepsilon_t\}$ i.i.d. standard Gaussian. This illustrates the behavior of the procedures under misspecification. More precisely, we use the extreme-value statistic as in Theorem 3.1 c), where we truncate the simulations at $10m$.



(a) $m = 100$ (b) $m = 200$ (c) $m = 300$

Figure 6.3: Truncated scaled density estimate of the run length for the misspecified neural network monitoring with a true underlying AR(1)-model, $\theta_0 = (1, 0.3)$

Table 6.2 shows the results of a simulation study based on 1000 repetitions. While the size for a moderate historical length is slightly liberal it becomes conservative for longer historic data sets. This effect is probably due to misspecification where a somewhat longer training sample is needed in order to get an approximation that also holds for future observations. The power is good but as expected depends on the kind of parameter change present. This becomes even clearer when looking at the scaled density estimate of the run length as given in Figure 6.3. Due to early detection and a long observation period, we truncate the scaled run length density estimate at the point where it is getting close to zero.

Data Example

In this section we consider the daily closing value Y_t of the SAP stock (for the period January 2007 to April 2011), for which the log-returns are defined as $R_t = 100(\log(Y_t) - \log(Y_{t-1}))$. The log returns can be modeled, e.g., by a β-ARCH models and in particular a first order autoregressive conditional heteroscedastic model -ARCH(1)-, i.e.,

$$R_t = \eta_t \sqrt{\omega + \alpha R_{t-1}^2}, \quad \omega > 0 \text{ and } 0 \leq \alpha < 1,$$

where the η_t are i.i.d. with mean zero and finite variance. Hence,

$$\log(R_t^2) = \log(\omega + \alpha R_{t-1}^2) + \log(\eta_t^2)$$

can be regarded as a first order nonlinear parametric autoregressive process. However, in practice some of the observed squared log-returns are close or equal to zero, as one can see in Figure 6.4. Therefore, instead of modeling squared log-return, we follow Fuller [20] and suggest

$$X_t = \log\left(R_t^2 + \iota\hat{\sigma}_m^2\right) - \frac{\iota\hat{\sigma}_m^2}{R_t^2 + \iota\hat{\sigma}_m^2} \text{ with, e.g., } \iota = 0.02,$$

and $\hat{\sigma}_m^2$ the empirical variance of the log-returns computed using the first m observations.

Additionally, we assume that X_t can be modeled as a first order nonlinear and nonparametric autoregressive process, i.e.,

$$X_t = g(X_{t-1}) + \varepsilon_t$$

with ε_t i.i.d. zero mean and finite variance. Indeed, this is a special case of the model defined in equation (6.4).

Finally, we apply the detector based on the monitoring function in (6.8) with $\widehat{\mathbf{A}}_H$ (and a neural network as in (6.5) with 2 hidden neurons) to the latter model.

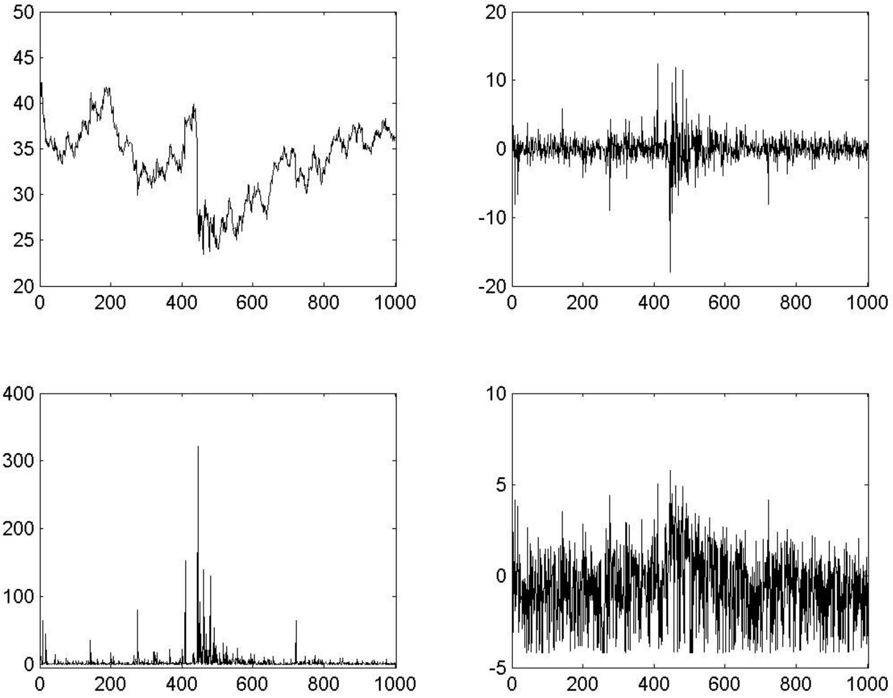


Figure 6.4: Upper panel: SAP stock and log-returns, lower panel: squared log-returns, Fuller type transformation of the squared-returns (July 2007 to June 2011)

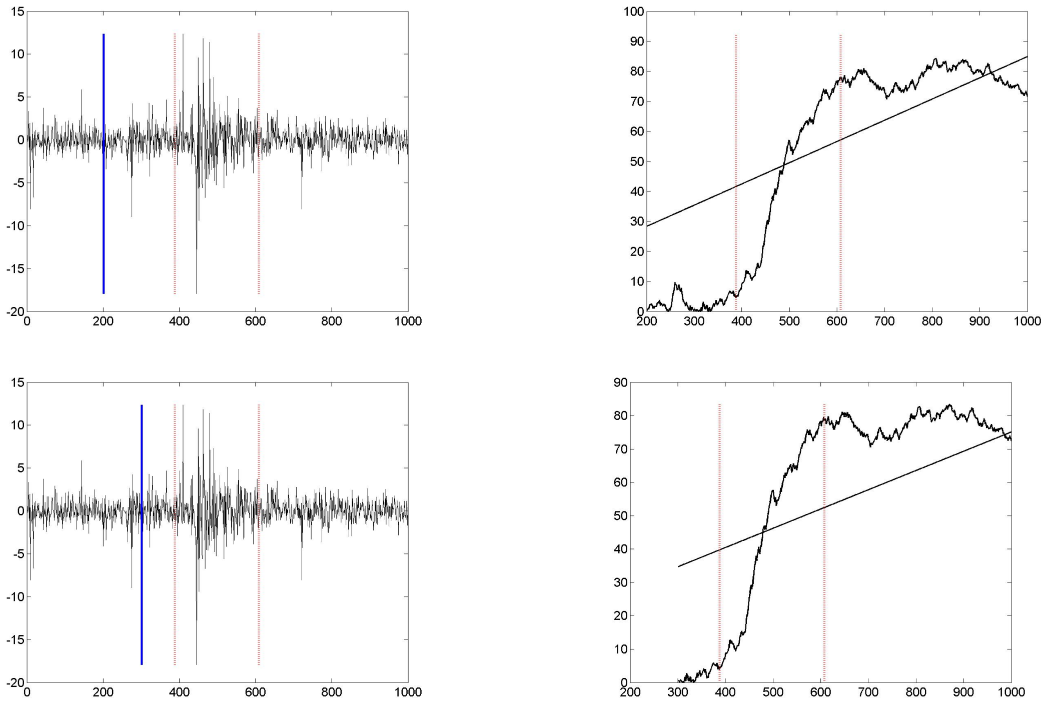


Figure 6.5: Detector for SAP data (nominal 5% level) Left: Data, where the solid line indicates the split between historic and monitoring data. Right: Corresponding detectors. The dotted lines indicate estimated change points (based on an offline procedure on the full data set)

Using the offline procedure of Kirch and Tadjuidje Kamgaing [30] in combination with a binary segmentation step yields two possible break points at 26.06.2008 as well as 21.05.2009 probably associated with the financial crises. In Figure 6.5 the Fuller type transformed log-returns as well as the monitoring chart are given, where the solid line indicates where the historic data ends and the monitoring starts and the dotted lines give the two possible change points. Obviously, the sequential procedure raises an alarm quickly after the first possible change point agreeing with the offline procedure.

6.4 Binary models

Binary time series are important in applications, where one is observing whether a certain event has or has not occurred within a given time frame. Wilks and Wilby [50] for example observe, whether it has been raining on a specific day, Kauppi and Saikkonen [27] and Startz [45] observe whether or not a recession has occurred in a given month. A common binary time series model is given by

$$X_t \mid X_{t-1}, X_{t-2}, \ldots, Z_{t-1}, Z_{t-2}, \ldots \sim \mathrm{Bern}(\pi_t(\boldsymbol{\beta})), \text{ with } g(\pi_t(\boldsymbol{\beta})) = \boldsymbol{\beta}^T \mathbb{Z}_{t-1}, \quad (6.9)$$

for a regressor $\mathbb{Z}_{t-1} = (Z_{t-1}, \ldots, Z_{t-p})^T$, which can be purely exogenous, purely autoregressive or a mixture of both. Typically, the canonical link function $g(x) = \log(x/(1-x))$

is used and statistical inference is based on the partial likelihood scores, which are defined by the following estimation function

$$G((X_t, \mathbb{Z}_{t-1}), \boldsymbol{\beta}) = \mathbb{Z}_{t-1}(X_t - \pi_t(\boldsymbol{\beta})) \tag{6.10}$$

for the canonical link function above.

The moment conditions in Assumptions B.2 and B.3 (for $H = G$) are fulfilled if $\mathbb{Z}_t$ has third moments, B.2 d) and A.3 follow immediately if $(X_t, \mathbb{Z}_{t-1})$ is strong mixing with exponential rates. For $\mathbb{Z}_{t-1} = (X_{t-1}, \ldots, X_{t-p})^T$, i.e. the standard binary autoregressive model (BAR(p)), Wang and Li [48] showed the geometric ergodicity property which in turn implies strong mixing with exponential rates. In the general setup, considering some regularity assumptions on the exogenous process, one can prove that $(X_t, \ldots, X_{t-p+1}, Z_t, \ldots, Z_{t-q})$ is a Feller chain, for which Theorem 1 of Feigin and Tweedie [14] can be applied to derive its geometric ergodic property (see Kirch and Tadjuidje Kamgaing [31] for details on this issue). Alternatively, invariance principles based on results of Eberlein [13] can be used (for details we refer to Fokianos et al. [16], Proposition 1).

For more general alternatives of the type considered in B.3 the moment conditions reduce to the existence of third moments of the regressor after the change by the boundedness of X_t^* and $\pi_t(\boldsymbol{\beta})$, while the same arguments as in the proof of Theorem 1.3.2 in Kirch and Tajduidje Kamgaing [33] give B.5 b). If both X_t and Y_t^* follow BAR-models, all parameter changes are asymptotically detected by the identifiability of the parameter via the partial score function.

Simulation study

We will now illustrate the monitoring for binary data by using a first order binary autoregressive process as in (6.9) with $\mathbb{Z}_{t-1} = (1, X_{t-1})$. We use the closed-end monitoring statistic with $H = G$ as in (6.10) with $N = 5$ and $w(m,k) = m^{-1/2}\left(1 + \frac{k}{m}\right)^{-1}$. We can then consistently estimate $\boldsymbol{S}_1 = \boldsymbol{S}_2 = \mathbb{E}\mathbb{Z}_{t-1}\mathbb{Z}_{t-1}^T \pi_t(\hat{\beta}_0)(1 - \pi_t(\hat{\beta}_0))$ by

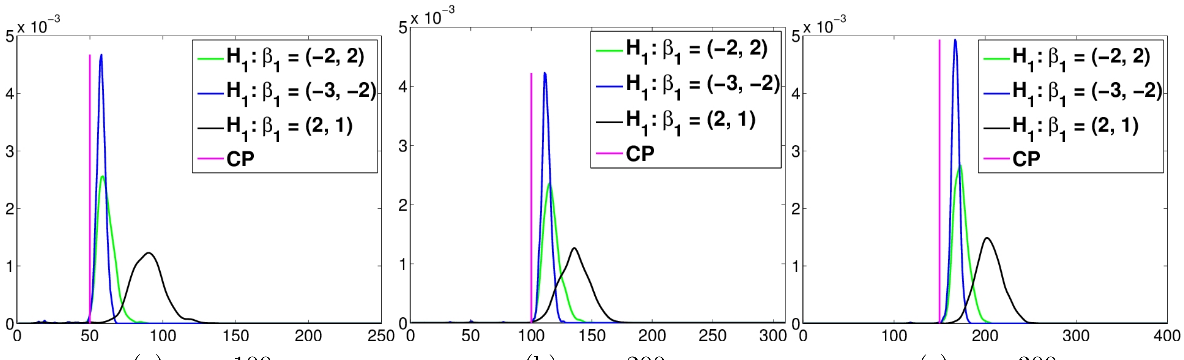
$$\widehat{\Sigma} = \frac{1}{m}\sum_{t=1}^{m} \mathbb{Z}_{t-1}\mathbb{Z}_{t-1}^T \pi_t(\widehat{\beta}_m)(1 - \pi_t(\widehat{\beta}_m)),$$

where $\widehat{\beta}_m$ is estimated based on the estimation function G and the historic data set only. Theorem 3.1 gives the null asymptotics in this situation.

	$H_0 : \beta_0 = (2, -2)$			$H_1 : \beta_1 = (-2, 2)$			$H_1 : \beta_1 = (-3, -2)$			$H_1 : \beta_1 = (2, 1)$		
m	100	200	300	100	200	300	100	200	300	100	200	300
	0.034	0.037	0.053	1	1	1	1	1	1	1	1	1

Table 6.3: Empirical size and power for BAR(1) monitoring at the nominal 5% level, β_0 denotes the null parameters before the change, β_1 the parameters after the change

Table 6.3 reports the empirical size and power (based on 1000 repetitions) for the nominal 5% level and various alternatives, where a change always occurred at time $\frac{m}{2}$ after the monitoring started.



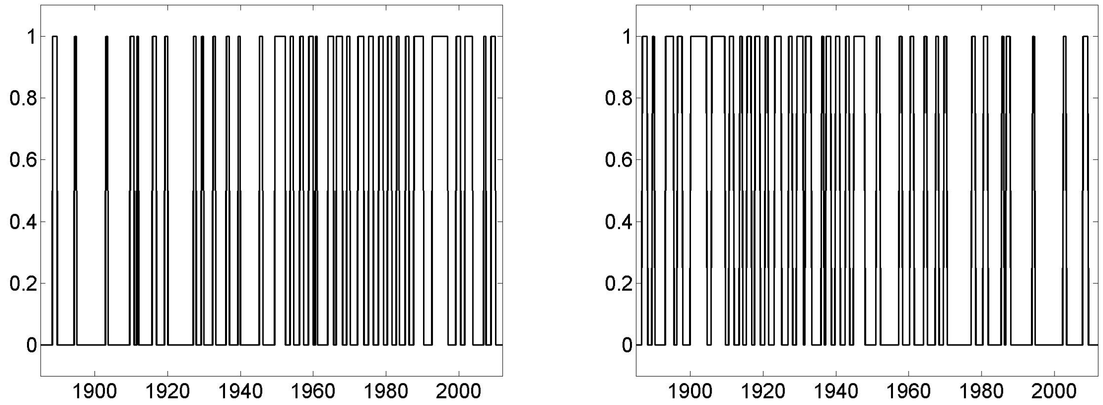
(a) $m = 100$ (b) $m = 200$ (c) $m = 300$

Figure 6.6: Truncated scaled density estimate of the run length for BAR(1)-model for the nominal 5% level, $\beta_0 = (2, -2)$

Figure 6.6 gives a truncated version of the scaled density estimator of the run length. The monitoring is conservative under the null hypothesis and detects the considered changes with empirical power one and relatively quickly after they occur.

Data Example

We apply the above test statistic to the US recession data (see Figure 6.7) for the period 1855–2012 for monthly data resp. for quarterly data.[1]



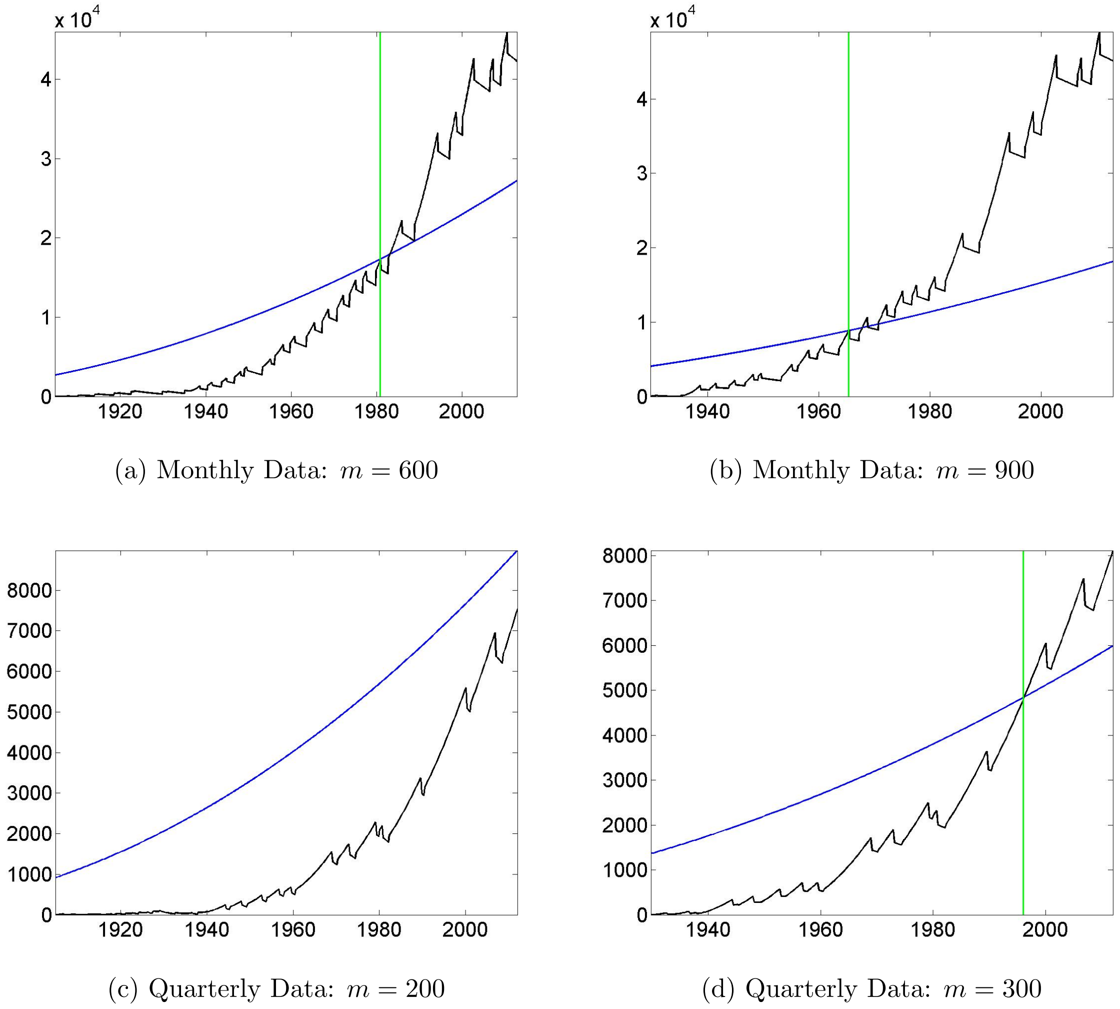
(a) Monthly Data (b) Quarterly Data

Figure 6.7: US recession data

[1] This data set can be downloaded from the National Bureau of Economic Research at http://research.stlouisfed.org/fred2/series/USREC

The quarterly version of this data set has been analyzed by Kirch and Tadjuidje Kamgaing [33] and Hudecová [42] in the context of offline change point detection. Their findings indicate the existence of a change point in the 1930s.

We use several historical data sets, where we check the non contamination assumption using an offline testing procedure (see, e.g. Kirch and Tadjuidje Kamgaing [33]). The corresponding detectors can be found in Figure 6.8.



(a) Monthly Data: $m = 600$

(b) Monthly Data: $m = 900$

(c) Quarterly Data: $m = 200$

(d) Quarterly Data: $m = 300$

Figure 6.8: Detectors for US recession data at nominal 5% level. The vertical lines indicate the stopping time.

Three out of the detector schemes (corresponding to different historical data sets) have detected the change point. For the fourth a clear increase in the detector is visible but it is has not yet rejected. The detection delay is shorter the later we start monitoring due to a relatively late change point. While this is typical in general, it seems to be particularly problematic in this situation of binary data. Using different detector sums with a shorter memory can significantly decrease the detection delay but has not yet been discussed in this general framework.

6.5 Poisson autoregressive time series

Another popular model for time series of counts is given by the Poisson autoregression, where we observe $X_1, \ldots, X_n$ with

$$X_t \mid X_{t-1}, \ldots, X_{t-p} \sim \text{Pois}(\lambda_t), \quad \lambda_t = f_\theta(\mathbb{X}_{t-1}), \quad \mathbb{X}_{t-1} = (X_{t-1}, \ldots, X_{t-p})^T. \tag{6.11}$$

If $f_\theta(\mathbf{x})$ is Lipschitz-continuous in $\mathbf{x}$ for all $\theta \in \Theta$ with Lipschitz constant strictly smaller than 1, then there exists a stationary ergodic solution of the (6.11) which is β-mixing with exponential rate (confer Neumann [40]). From this we obtain Assumption A.3.

Under suitable smoothness assumptions on f_θ in connection with suitable moment assumptions, one can derive the regularity conditions B.1 – B.5 for the least squares estimating functions in an analogous fashion as in Section 6.3 for the neural network function above. This is the approach taken by Franke et al. [17] in an offline setting.

We will now take a closer look at the INARCH(1)-model given by $\lambda_t = \omega + \alpha X_{t-1}$ with $0 < \delta,\ 0 < \delta \leq \omega \leq \Delta,\ 0 \leq \alpha \leq 1 - \delta < 1$ and the estimating function obtained from the partial log likelihood scores, i.e.

$$\sum_{t=1}^{n} \begin{pmatrix} 1 \\ X_{t-1} \end{pmatrix} \frac{(X_t - \lambda_t)}{\lambda_t} = \sum_{t=1}^{n} G((X_t, X_{t-1}), \theta).$$

Considering the Euclidean norm on $\mathbb{R}^2$ and using the compactness assumption on Θ, it follows,

$$\|G((X_t, X_{t-1}), \theta)\|^2 = (X_t - \lambda_t)^2 \left(\frac{1}{\lambda_t^2} + \frac{X_{t-1}^2}{\lambda_t^2} \right) \leqslant (X_t - \lambda_t)^2 \left(\frac{1 + X_{t-1}^2}{\delta^2} \right)$$
$$\leqslant \frac{1}{\delta^2} (X_t + X_{t-1} + \Delta)^2 (1 + X_{t-1}^2)$$

for all $\theta \in \Theta$. Therefore,

$$\mathbb{E} \sup_{\theta \in \Theta} \|G((X_t, X_{t-1}), \theta)\| \leq \frac{1}{\delta} \mathbb{E}(X_t + X_{t-1} + \Delta)(1 + X_{t-1}),$$

which is finite since the second moment of the process X_t exists (cf. Ferland et al. [15]), implying B. In fact, in [15], Proposition 6 even proves the finiteness of all moments for such a process.

The gradient of the estimating function is given by

$$\nabla G((X_t, X_{t-1}), \theta) = \begin{pmatrix} \frac{-X_t}{\lambda_t^2} & \frac{-X_t X_{t-1}}{\lambda_t^2} \\ \frac{-X_t X_{t-1}}{\lambda_t^2} & \frac{-X_t X_{t-1}^2}{\lambda_t^2} \end{pmatrix}.$$

Similarly, using the Euclidean norm on $\mathbb{R}^{2\times 2}$, it follows,

$$\|\nabla G((X_t, X_{t-1}), \theta)\|^2 = \frac{1}{\lambda_t^2} \sqrt{X_t^2 + 2X_t^2 X_{t-1}^2 + X_t^2 X_{t-1}^4} = \frac{X_t^2}{\lambda_t^2} (1 + X_{t-1}^2)^2$$
$$\leqslant \frac{X_t^2}{\delta^2} (1 + X_{t-1}^2)^2$$

for all $\theta \in \Theta$. Therefore,

$$\mathbb{E} \sup_{\theta\in\Theta} \|\nabla G((X_t, X_{t-1}), \theta)\| \leq \frac{1}{\delta} \mathbb{E} X_t (1 + X_{t-1}^2),$$

which is finite since third moments exist. Similarly,

$$\mathbb{E} \sup_{\theta\in\Theta} \|\nabla^2 G((X_t, X_{t-1}), \theta)\| < \infty$$

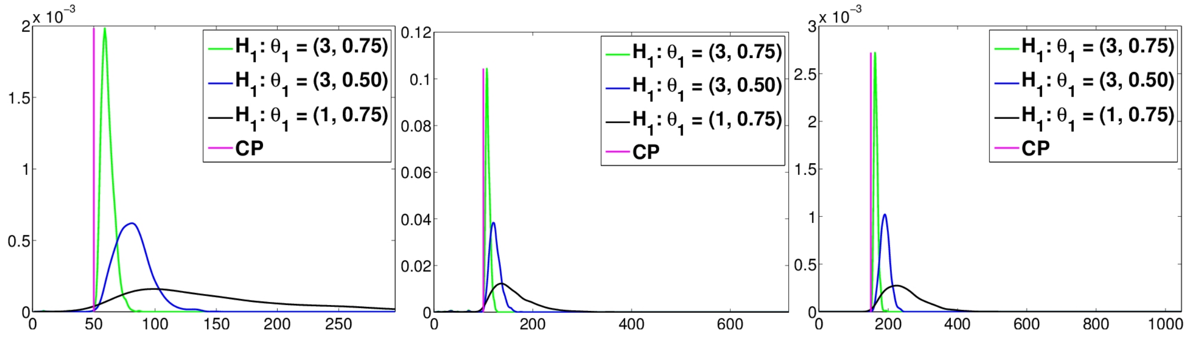
since the fourth moments exist, yielding B.3 for $H = G$.

Simulation study

In the simulations we consider a Poisson autoregressive model as in (6.11) with $\lambda_t = \theta_1 + \theta_2 X_{t-1}$. We use the closed-end monitoring procedure with $G = H$ as above, $w(m,k) = m^{-1/2}\left(1 + \frac{k}{m}\right)^{-1}$ and $N = 5$. We can then consistently estimate $\boldsymbol{S}_1 = \boldsymbol{S}_2 = \mathbb{E}\mathbb{Z}_{t-1}\mathbb{Z}_{t-1}^T \frac{(X_t-\lambda_t)^2}{\lambda_t^2}$ by the empirical covariance matrix of $\{G((X_t, X_{t-1}), \hat{\theta})\}$. Theorem 3.1 gives the null asymptotics in this situation.

	$H_0 : \theta_0 = (1, 0.5)$			$H_1 : \theta_1 = (3, 0.75)$			$H_1 : \theta_1 = (3, 0.5)$			$H_1 : \theta_2 = (1, 0.75)$		
m	100	200	300	100	200	300	100	200	300	100	200	300
	0.022	0.033	0.036	1	1	1	1	1	1	0.999	1	1

Table 6.4: Empirical size and power for the INARCH(1) model (at nominal 5% level) θ_0 indicates the null parameters, θ_1 the parameters after the change



(a) $m = 100$ (b) $m = 200$ (c) $m = 300$

Figure 6.9: Truncated scaled density estimate of the run length for INARCH(1)-model (at nominal 5% level), where the vertical lines indicate the change point, $\theta_0 = (1, 0.5)$

Table 6.4 shows the empirical size and power at the nominal 5% level for various alternatives, while Figure 6.9 shows a truncated version of scaled density of the run length. The test is clearly conservative under the null hypothesis but rejects all simulated alternatives with only one exception. The detection delay differs for different alternatives, where the smallest change has the longest detection delay.

Data Analysis

We will demonstrate the above methodology using the number of transactions per minute for the stock Ericsson B during July 3rd 2002 (confer Figure 6.10), where we do not take the first 5 and last 15 minutes of transaction time into account.

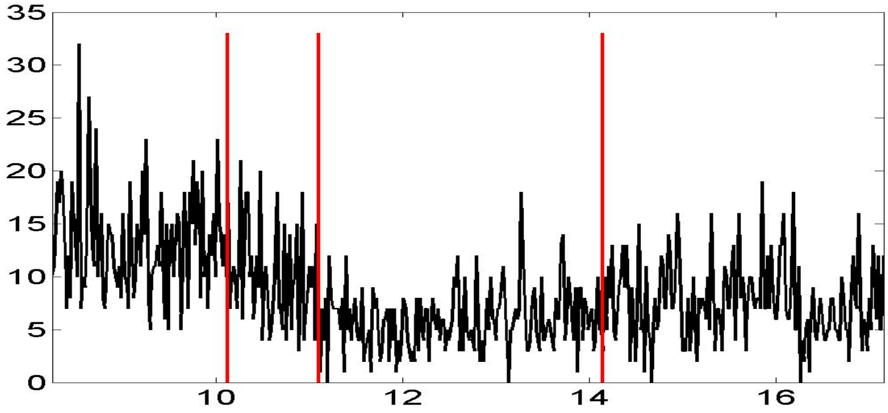
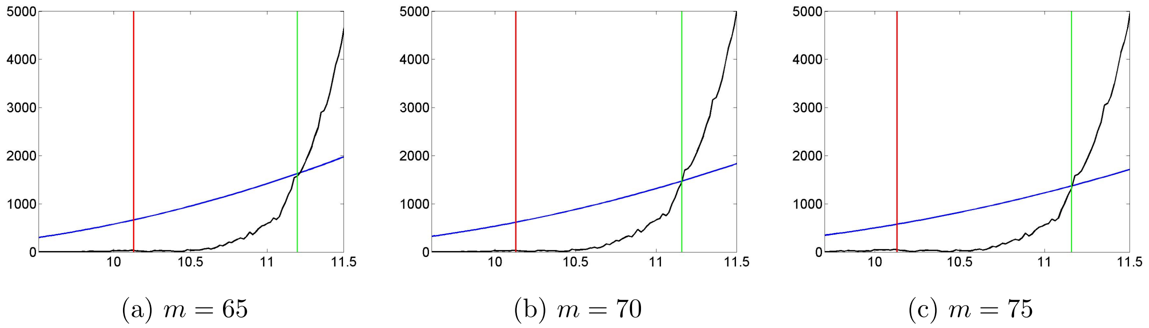


Figure 6.10: Stock Ericsson B: July 3rd 2002. The vertical red lines indicate estimated change points (by binary segmentation on the full data set)

Kirch and Tadjuidje [33] have analyzed this data set in an offline change setup. Their analysis indicates three possible change points using binary segmentation, which are given by the red vertical lines in the plot. In Figure 6.11 the detectors are given for various choices of m (all before the first change point indicated by the a posteriori analysis).



(a) $m = 65$

(b) $m = 70$

(c) $m = 75$

Figure 6.11: Value of the detector of change for the number of transactions per minute of the Ericson B stock data with different start dates for the monitoring. The vertical green lines indicate the estimated change points.

6.6 Conclusions

The above examples show that by varying the estimating function as well as the monitoring functions, we are able to tune the sequential change point detectors to many different situations. This ranges from detecting parameter changes in time series models, where misspecification can be taken into account, to developing detection procedures that are robust with respect to outliers. While the theory derived in the previous sections relies on asymptotic arguments, several simulations and data examples indicate that the procedures also work well in small samples.

7 Proofs

Proof of Theorem 3.1. For any symmetric positive semi-definite matrix $\mathbf{A}$, let us introduce the notation $\|Z\|_{\mathbf{A}}^2 = Z^T\mathbf{A}Z$, then by Assumption A.2 it holds for general weight functions $w(m,k)$

$$\begin{aligned}&\sup_{1\leqslant k<N(m)} w^2(m,k)\,\|\boldsymbol{S}(m,k)\|_{\boldsymbol{A}}^2\\ &=\sup_{1\leqslant k<N(m)} w^2(m,k)\left\|\sum_{j=m+1}^{m+k} H(\mathbf{X}_j,\theta_0)-\frac{k}{m}\boldsymbol{B}(\theta_0)\sum_{j=1}^{m}G(\mathbf{X}_j,\theta_0)\right\|_{\boldsymbol{A}}^2+o_P(1).\end{aligned}$$

For $\tilde{w}(m,k)=\rho(k/m)$ with ρ bounded, we conclude from the functional central limit theorem in A.3 (i), that for any $N>0$

$$\begin{aligned}&\sup_{1\leqslant k<Nm}\tilde{w}^2(m,k)\,\|\boldsymbol{S}(m,k)\|_{\boldsymbol{A}}^2\\ &=\sup_{j=1,\dots,p+1}\sup_{t\in I_j}\rho^2(t)\left\|\frac{1}{\sqrt{m}}\sum_{j=m+1}^{m+\lfloor mt\rfloor}H(\mathbf{X}_j,\theta_0)-\frac{\lfloor mt\rfloor}{m}\frac{1}{\sqrt{m}}\boldsymbol{B}(\theta_0)\sum_{j=1}^{m}G(\mathbf{X}_j,\theta_0)\right\|_{\boldsymbol{A}}^2+o_P(1)\\ &\xrightarrow{\mathcal{D}}\sup_{0\leqslant t\leqslant N}\rho^2(t)\,\|\mathbf{W}_1(1+t)-\mathbf{W}_1(1)-t\mathbf{W}_2(1)\|_{\boldsymbol{A}}^2\,.\end{aligned}$$

Noting that $\{\mathbf{W}_1(1+t)-\mathbf{W}_1(t) : t\geqslant 0\}$ is again a Wiener process with covariance matrix $\mathbf{\Sigma}_1$ independent of $\mathbf{W}_2(1)$ yields the first assertion in a). For the more general weight functions $\tilde{w}(m,k)$ as in Assumption A.1 (a), analogous arguments show the convergence for $k\geqslant\tau m$ towards $\sup_{\tau\leqslant t\leqslant N}\rho^2(t)\,\|\mathbf{W}_1(1+t)-\mathbf{W}_1(1)-t\mathbf{W}_2(1)\|_{\boldsymbol{A}}^2$ for any $\tau>0$. By Assumptions A.1 (a) as well as A.3 (a) (i) and (ii) it holds for some generic constant $C>0$ and $\gamma<\alpha<1/2$

$$\begin{aligned}&\sup_{1\leqslant k<\tau m} w^2(m,k)\left\|\sum_{j=m+1}^{m+k}H(\mathbf{X}_j,\theta_0)-\frac{k}{m}\boldsymbol{B}(\theta_0)\sum_{j=1}^{m}G(\mathbf{X}_j,\theta_0)\right\|_{\boldsymbol{A}}^2\\ &\leqslant C\sup_{0\leqslant t<\tau}t^{2\alpha}\rho^2(t)\\ &\quad\cdot\left(\sup_{1/m\leqslant t\leqslant\tau}\frac{1}{m^{1-2\alpha}k^{2\alpha}}\left\|\sum_{j=m+1}^{m+\lfloor mt\rfloor}H(\boldsymbol{X}_j,\theta_0)\right\|^2+\tau^{1-2\alpha}\left\|\frac{1}{\sqrt{m}}\boldsymbol{B}(\theta_0)\sum_{j=1}^{m}G(\mathbf{X}_j,\theta_0)\right\|^2\right)\\ &\xrightarrow{P}0\end{aligned}\tag{7.1}$$

as $\tau \to 0$ uniformly in m. An analogous assertion can be obtained for the limiting Wiener processes concluding a).

Analogously, we get by Assumption A.1 (b) and A.3 (a)(ii) that

$$\sup_{k \geqslant Tm} w^2(m,k) \left\| \sum_{j=m+1}^{m+k} H(\mathbf{X}_j, \theta_0) \right\|_{\boldsymbol{A}}^2 \xrightarrow{P} 0 \tag{7.2}$$

as $T \to \infty$ uniformly in m as well as an analogous expression for the limiting Wiener process. From the functional central limit theorem in A.3 (a) (i) and A.1 (b) it follows for any $0 < \tau < T < \infty$ that

$$\sup_{k \geqslant \tau m} \left\| w(m, \min(k, Tm)) \sum_{j=m+1}^{m+\min(k,mT)} H(\mathbf{X}_j, \theta_0) - w(m,k)\, \frac{k}{m} \boldsymbol{B}(\theta_0) \sum_{j=1}^{m} G(\mathbf{X}_j, \theta_0) \right\|_{\boldsymbol{A}}^2$$

$$\xrightarrow{\mathcal{D}} \sup_{t \geqslant \tau} \left\| \rho(\min(t,T))(\mathbf{W}_1(1+t) - \mathbf{W}_1(1)) - t\rho(t)\mathbf{W}_2(1) \right\|_{\boldsymbol{A}}. \tag{7.3}$$

Carefully combining (7.1) –(7.3) yields b).

The proof of c) is analogous to Horváth et al. [24], proof of Theorem 1.1, but in the multivariate setting. The only difference occurs in (3.12), where we prove instead that (with the notation of that paper)

$$\lim_{m\to\infty} P\left(a(\log m) \sup_{\frac{a(m)}{m+a(m)} \leqslant s \leqslant \frac{c}{1+c}} \frac{\sqrt{\sum_{j=1}^d W_j^2(s)}}{\sqrt{s}} - b_d(\log m) \leqslant t \right) = \exp(-e^{-t}), \tag{7.4}$$

for independent Wiener processes $\{W_j(\cdot)\}$. First note, that

$$\sup_{a(m)/(m+a(m)) \leqslant s \leqslant 1} \frac{\sqrt{\sum_{j=1}^d W_j^2(s)}}{\sqrt{s}} = \sup_{1 \leqslant t \leqslant (m+a(m))/a(m)} \frac{\sqrt{\sum_{j=1}^d W_j^2(1/t)}}{\sqrt{1/t}}$$
$$\overset{\mathcal{D}}{=} \sup_{1 \leqslant t \leqslant (m+a(m))/a(m)} \frac{\sqrt{\sum_{j=1}^d W_j^2(t)}}{\sqrt{t}}.$$

By the proof of Lemma 2.2 in Horváth [22] we get (7.4) with $a(\log m)$ replaced by $a(\log((m+a(m))/a(m)))$ and $b_d(\log m)$ by $b_d(\log((m+a(m))/a(m)))$. Since

$$a(\log m)|a(\log m) - a(\log((m+a(m))/a(m)))| \to 0,$$
$$b_d(\log m) - b_d(\log((m+a(m))/a(m))) \to 0,$$

assertion (7.4) follows, completing the proof of c). ■

Proof of Theorem 3.2. Part a) can be found in Hušková and Koubková [26], proof of Theorem 2.1, confer also Horváth et al. [23] for the univariate case. Part b) proceeds analogously on noting that

$$\{\sigma_2^2(W_1(t) - tW_2(1)) : 0 \leqslant t < \infty\} \overset{\mathcal{D}}{=} \left\{ (\sigma_1^2 + \sigma_2^2 t)\, W\left(\frac{\sigma_2^2\, t}{\sigma_1^2 + \sigma_2^2 t} \right) : 0 \leqslant t < \infty \right\}$$

for a standard Wiener process $\{W(\cdot)\}$. ■

Proof of Theorem 4.1. For $\widetilde{k} > k^*$ it holds

$$\begin{aligned}\sum_{t=m+1}^{m+\tilde{k}} H(\mathbf{X}_t, \widehat{\theta}_m) &= \sum_{t=m+1}^{m+k^*} H(\mathbf{X}_t, \widehat{\theta}_m) + \sum_{t=m+k^*}^{m+\tilde{k}} H(\mathbf{X}_t, \widehat{\theta}_m) \\ &=: \mathbf{S}_{H_0}(m, k^*) + \mathbf{S}_{H_1}(m + k^*, \tilde{k}).\end{aligned}$$

Under Assumption A.4 a) and b) an application of Theorem 3.1 implies

$$\frac{1}{m}\mathbf{S}_{H_0}(m, k^*) = o_P(1),$$

while (4.1) implies for $\tilde{k} = \lfloor m x_0 \rfloor$

$$\frac{1}{m}\mathbf{S}_{H_1}(m + k^*, \tilde{k}) = (x_0 - \vartheta)\, \mathbf{E}_H + o_P(1).$$

Together this yields by an application of the Cauchy-Schwarz inequality for $\mathbf{E}_H^T \mathbf{A} \mathbf{E}_H \neq 0$

$$\max_{k \geqslant 1} w^2(m, k)\mathbf{S}(m, k)^T \mathbf{A}\mathbf{S}(m, k) \geqslant m\rho^2 \left(x_0 + o(1)\right)(x_0 - \vartheta)^2 \left(\mathbf{E}_H^T \mathbf{A}\mathbf{E}_H + o_P(1)\right) \xrightarrow{P} \infty,$$

showing that the corresponding test has asymptotic power one.

For the open-end procedure with $k^* = O(m)$ analogous arguments give the assertion, for $k^*/m \to \infty$ consider $\widetilde{k} = 2k^*$ and note that by Theorem 3.1

$$\frac{1}{k^*}\mathbf{S}_{H_0}(m, k^*) = o_P(1)$$

and by (4.2)

$$\frac{1}{k^*}\mathbf{S}_{H_1}(m + k^*, \tilde{k}) = \mathbf{E}_H + o_P(1),$$

which implies

$$\max_{k \geqslant 1} w^2(m, k)\mathbf{S}(m, k)^T \mathbf{A}\mathbf{S}(m, k) \geqslant \frac{\tilde{k}^2}{4m}\rho^2 \left(\frac{\tilde{k}}{m}\right) \left(\mathbf{E}_H^T \mathbf{A}\mathbf{E}_H + o_P(1)\right) \xrightarrow{P} \infty,$$

proving that the open-end procedure as in Theorem 3.1 b) has asymptotic power one. Similarly, one can show for the statistic in Theorem 3.1 c) that

$$\frac{1}{\sqrt{\log\log m}} \sup_{1 \leqslant k < \infty} \frac{\sqrt{\boldsymbol{S}(m, k)^T \mathbf{A}\boldsymbol{S}(m, k)}}{\sqrt{m}\left(1 + \frac{k}{m}\right)\left(\frac{k}{m+k}\right)^{1/2}} \xrightarrow{P} \infty,$$

implying that the corresponding statistic has asymptotic power one. ■

Proof of Proposition 5.1. The proof of a) follows analogously to the proof of Proposition 1.2.1 in Kirch and Tajduidje Kamgaing [33], the proof of b) is analogous to Theorem 3 in Kirch and Tadjuidje Kamgaing [30]. ■

Proof of Proposition 5.2. By definition of $\widehat{\theta}_m$ it holds

$$\sum_{t=1}^{m} G\left(\mathbf{X}_t, \widehat{\theta}_m\right) = 0. \tag{7.5}$$

From this we can conclude

$$\begin{aligned}
&\sum_{t=m+1}^{m+k} H\left(\mathbf{X}_t, \widehat{\theta}_m\right) - \left(\sum_{i=m+1}^{m+k} H\left(\mathbf{X}_t, \theta_0\right) - \frac{k}{m}\boldsymbol{B}(\theta_0)\sum_{t=1}^{m} G\left(\boldsymbol{X}_t, \theta_0\right)\right) \\
&= \sum_{t=m+1}^{m+k} \left(H\left(\mathbf{X}_t, \widehat{\theta}_m\right) - H\left(\mathbf{X}_t, \theta_0\right)\right) - \frac{k}{m}\sum_{t=1}^{m}\left(\boldsymbol{B}(\theta_0)\, G\left(\mathbf{X}_t, \widehat{\theta}_m\right) - \boldsymbol{B}(\theta_0)\, G\left(\mathbf{X}_t, \theta_0\right)\right) \\
&=: D_1(m,k) - D_2(m,k).
\end{aligned}$$

Let H_j denote the j-th component function of H, then a Taylor expansion yields

$$H_j(\mathbf{X}_t, \widehat{\theta}_m) - H_j(\mathbf{X}_t, \theta_0) = \nabla H_j(\mathbf{X}_t, \theta_0)^T(\widehat{\theta}_m - \theta_0) + \frac{1}{2}(\widehat{\theta}_m - \theta_0)^T \nabla^2 H_j(\mathbf{X}_t, \xi_j)(\widehat{\theta}_m - \theta_0), \tag{7.6}$$

where $\nabla H_j(\boldsymbol{X}_t, \theta)$ is the gradient with respect to θ and $\nabla^2 H_j(\boldsymbol{X}_t, \theta)$ is the Hessian matrix, ξ_j is between θ_0 and $\widehat{\theta}_m$ element wise. By assumption B.1 and B.3 and a uniform law of large numbers for stationary and ergodic processes (cf. Ranga Rao [41], Theorem 6.5) it holds

$$\sup_{k\geqslant 1}\sup_{\xi\in\Theta}\frac{1}{k}\sum_{t=m+1}^{m+k}\|\nabla^2 H_j(\mathbf{X}_t, \xi)\|_\infty = O_P(1),$$

where $\|(\alpha_{i,j})\|_\infty = \max_{i,j}|\alpha_{i,j}|$. Together with (7.6) this yields uniformly in k

$$\sum_{t=m+1}^{m+k}(H_j(\mathbf{X}_t, \widehat{\theta}_m) - H_j(\mathbf{X}_t, \theta_0)) = \sum_{t=m+1}^{m+k}\nabla H_j(\mathbf{X}_t, \theta_0)^T(\widehat{\theta}_m - \theta_0) + O_P\left(k\|\widehat{\theta}_m - \theta_0\|^2\right). \tag{7.7}$$

An application of the ergodic theorem yields

$$\frac{1}{l}\sum_{t=1}^{l}(\nabla H_j(\mathbf{X}_t, \theta_0)^T - \mathbb{E}\nabla H_j(\mathbf{X}_t, \theta_0)^T) = o(1) \quad a.s. \qquad (l \to \infty). \tag{7.8}$$

Conditions A.1 imply that for some $C > 0$ and some $0 \leqslant \gamma < 1/2$

$$w(m,k) \leqslant \begin{cases} Cm^{\gamma-1/2}k^{-\gamma}, & k \leqslant m, \\ Cm^{1/2}\, k^{-1}, & k > m. \end{cases} \tag{7.9}$$

Proposition 5.1 b) together with (7.6) – (7.9) yields (as $m \to \infty$)

$$\begin{aligned}
&\sup_{k\geq 1} w(m,k)\,\|D_1(m,k) - k\mathbb{E}\nabla H(\mathbf{X}_1,\theta_0)^T(\widehat{\theta}_m - \theta_0)\| \\
&= O_P(1) \sup_{k\leqslant\sqrt{m}} \left(\frac{k}{m}\right)^{1-\gamma} + o_P(1) \sup_{\sqrt{m}<k\leqslant m} \left(\frac{k}{m}\right)^{1-\gamma} + o_P(1) = o_P(1). \qquad (7.10)
\end{aligned}$$

Analogously

$$\sup_{k\geqslant 1} w(m,k)\,\|D_2(m,k) - k\mathbb{E}\nabla \boldsymbol{B}(\theta_0)\,G(\mathbf{X}_1,\theta_0)^T(\widehat{\theta}_m - \theta_0)\| = o_P(1). \qquad (7.11)$$

Since by definition of $\boldsymbol{B}(\theta_0)$ it holds

$$\mathbb{E}\nabla H = \boldsymbol{B}(\theta_0)\mathbb{E}\nabla G = \mathbb{E}\nabla\boldsymbol{B}(\theta_0)G,$$

the assertion follows. ■

Proof of Proposition 5.3. The assertions follow similarly to the proof of Proposition 5.2 by a Taylor expansion in connection with a (uniform) ergodic theorem. ■

Proof of Proposition 5.4. By the assumption and an application of the Cauchy-Schwarz inequality it holds

$$\begin{aligned}
&\frac{1}{l}\sum_{j=m+k^*+1}^{m+k^*+l} \|H(\boldsymbol{X}_j,\widehat{\theta}_m) - H(\boldsymbol{X}_j^*,\widehat{\theta}_m)\| \\
&\leqslant C\frac{1}{l}\sum_{j=m+k^*+1}^{m+k^*+l}\|\boldsymbol{R}_j\|^2 + \sqrt{\frac{1}{l}\sum_{j=m+k^*+1}^{m+k^*+l}\|\boldsymbol{R}_j\|^2\,\frac{1}{l}\sum_{j=m+k^*+1}^{m+k^*+l}\|F(\boldsymbol{X}_j^*)\|^2} = o_P(1)
\end{aligned}$$

by an application of the ergodic theorem. The assertion then follows from Proposition 5.3. ■

Proof of Proposition 5.5. Let $\{\boldsymbol{X}_t^*\}$ be a stationary Markov chain with the same transition kernels as $\{\boldsymbol{X}_t\}$ with starting value $\mathbf{x}_0^*$ from the stationary distribution. By Theorem 21.12 of Lindvall [38] it holds $X_t = X_t^*$ for all $t > T$, where T is an almost surely finite random time. From this it follows that $\max_{t\geqslant 1}\|\boldsymbol{X}_t - \boldsymbol{X}_t^*\|^2$ is almost surely bounded so that the assertion follows for $\boldsymbol{R}_t = \boldsymbol{X}_t - \boldsymbol{X}_t^*$. ■

Acknowledgments

The work was supported by DFG grant KI 1443/2-2. The work was done while the first author was at the Karlsruhe Institute of Technology (KIT), where her position was financed by the Stifterverband für die Deutsche Wissenschaft by funds of the Claussen-Simon-trust. We would like to thank Roland Fried for asking the question whether it is in principle possible to use different estimating functions for the estimation and the monitoring which lead to a yet another generalization in this paper.

References

[1] Anderson, T. W. A modification of the sequential probability ratio test to reduce the sample size. *The Annals of Mathematical Statistics*, 31:165–197, 1960.

[2] Andreou, E., and Ghysels, E. Monitoring disruptions in financial markets. *J. Econometrics*, 135:77–124, 2006.

[3] Aston, J. and Kirch, C. Change-points in high-dimensional settings. 2014. In preparation.

[4] Aue, A., Berkes, I., and Horváth, L. Strong approximation for the sums of squares of augmented GARCH sequences. *Bernoulli*, 2006. To appear.

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

[6] Aue, A. and Horváth, L. Delay time in sequential detection of change. *Stat. Probab. Lett.*, 67(3):221–231, 2004.

[7] Aue, A., Horváth , L.., Hušková , M., and Kokoska, P. Change-point monitoring in linear models. *Econometrics Journal*, 9:373–403, 2006.

[8] Aue, A., Horváth , L., and Reimherr, M.L. Delay times of sequential procedures for multiple time series regression models. *Journal of Econometrics*, 149(2):174 – 190, 2009.

[9] Berkes, I., Gombay, E., Horváth , L., and Kokoszka, P. Sequential change-point detection in GARCH(p,q) models. *Econometric theory*, 20, 2004.

[10] Chochola, O. *Robust Monitoring Procedures for Dependent Data.* PhD thesis, Charles University Prague, 2013.

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

[12] Ciuperca, G. Two tests for sequential detection of a change-point in a nonlinear model. *ournal of Statistical Planning and Inference*, 143(10):1719 – 1743, 2013.

[13] Eberlein, E. On strong invariance principles under dependence assumptions. *Ann. Probab.*, 14:260–270, 1986.

[14] Feigin, P. D. and Tweedie, R. L. Random coefficient autoregressive processes: A Markov chain analysis of stationarity and finiteness of moments. *Journal of Time Series Analysis*, 6(1):1–14, 1985.

[15] Ferland, R., Latour, A., and Oraichi, D. Integer-valued garch process. *Journal of Time Series Analysis*, 27(6), 1984.

[16] Fokianos, K., Gombay, E., and Hussein, A. Retrospective change detection for binary time series models. *Journal of Statistical Planning and Inference*, 145:102 – 112, 2014.

[17] Franke, J., Kirch, C., and Tadjuidje Kamgaing, J. . Changepoints in times series of counts. *J. Time Series Analysis*, 33:757–770, 2012.

[18] Franke, J. and Mabouba, D. Estimating market risk with neural networks. *Statistic Decision*, 30:63–82, 2006.

[19] Fried, R., and Imhoff, M. On the online detection of monotonic trends in time series. *Biom. J.*, 46:90–102, 2004.

[20] Fuller, W.A. *Introduction to Statistical Time Series.* Wiley, New York, 1996.

[21] Hall, A. R. *Generalized Method of Moments.* Advanced Texts in Econometrics Series. Oxford University Press, 2005.

[22] Horvath, L. The maximum likelihood method for testing changes in the parameters of normal observations. *The Annals of Statistics*, 21(2):pp. 671–680, 1993.

[23] Horváth, L., Hušková, M., Kokoszka, P., and Steinebach, J. Monitoring changes in linear models. *J. Statist. Plann. Inference*, 126:225–251, 2004.

[24] Horváth, L., Kokoszka, P., and Steinebach, J. Testing for changes in multivariate dependent observations with an application to temperature changes. *J. Multivariate Anal.*, 68:96–119, 1999.

[25] Horváth, L., Kokoszka, P., and Steinebach, J. On sequential detection of parameter changes in linear regression. *Stat. Probab. Lett.*, 77(9):885–895, 2007.

[26] Hušková, M. and Koubková, A. Monitoring jump changes in linear models. *J. Statist. Res.*, 39:51–70, 2005.

[27] Kauppi, H., and Saikkonen, P. Predicting US recessions with dynamic binary response models. *Review of Economics and Statistics*, 90:777–791, 2008.

[28] C. Kirch, B. Muhsal, and H. Ombao. Detection of changes in multivariate time series with application to eeg data. 2013. Preprint, Karlsruhe Institute of Technology.

[29] Kirch, C. and Tadjuidje Kamgaing, J. . An online approach to detecting changes in nonlinear autoregressive models. *preprint*, 2011.

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

[31] Kirch, C. and Tadjuidje Kamgaing, J. . Geometric ergodicity of binary autoregressive models with exogenous variables. *preprint*, 2013.

[32] Kirch, C. and Tadjuidje Kamgaing., J. . A uniform central limit theorem for neural network-based autoregressive processes with applications to change-point analysis. *Statistics*, 48:1187–1201, 2014.

[33] Kirch, C. and Tadjuidje Kamgaing, J. Detection of change points in discrete-valued time series. In R.A.Davis, S.A. Holan, R.B. Lund, and N. Ravishanker, editors, *Handbook of Discrete Valued Time series.* Springer Berlin Heidelberg, 2014+.

[34] Komlós, J., Major, P., and Tusnády, G. An approximation of partial sums of independent rvs and the sample df. i. *Z. Wahrsch. verw. Geb.*, 32:111–131, 1975.

[35] Komlós, J., Major, P., and Tusnády, G. An approximation of partial sums of independent rvs and the sample df. ii. *Z. Wahrsch. verw. Geb.*, 34:33–58, 1976.

[36] Koubková, A. *Sequential change-point analysis.* PhD thesis, Charles University Prague, 2006.

[37] Kuelbs, J., and Philipp, W. Almost sure invariance principles for partial sums of mixing b-valued random variables. *Ann. Probab.*, 8:1003–1036, 1980.

[38] Lindvall, T. *Lectures on the coupling method. Corrected reprint of the 1992 original.* Mineola, NY: Dover Publications, corrected reprint of the 1992 original edition, 2002.

[39] Meyn, S.P. and Tweedie, R.L. . *Markov Chains and Stochastic Stability.* Oxford university press, Oxford, 1990.

[40] Neumann, M. H. Absolute regularity and ergodicity of poisson count processes. *Bernoulli*, 17(4):1268–1284, 2011.

[41] Ranga Rao, R. Relation between weak and uniform convergence of measures with applications. *Ann. Math. Statist.*, 33:659–680, 1962.

[42] Hudecová , S. Structural changes in autoregressive models for binary time series. *Journal of Statistical Planning and Inference*, 143(10), 2013.

[43] Schmitz, A and Steinebach, J. A note on the monitoring of changes in linear models with dependent errors. In Paul Doukhan, Gabriel Lang, Donatas Surgailis, and Gilles Teyssire, editors, *Dependence in Probability and Statistics*, Lecture Notes in Statistics, pages 159–174. Springer Berlin Heidelberg, 2010.

[44] Siegmund, D. Repeated significance tests for a normal mean. *Biometrika*, 64:177–189, 1977.

[45] Startz, R. Binomial autoregressive moving average models with an application to us recession. *Journal of Business & Economic Statistics*, 26:1–8, 2008.

[46] Stockis, J.-P., Franke, J., and Tadjuidje Kamgaing, J. On geometric ergodicity of charme models. *J. Time Series Analysis*, 31:141–152, 2010.

[47] Tong, H. *Nonlinear time series: a dynamical system approach.* Springer, London, 1993.

[48] Wang, C. and Li, W. K. On the autopersistence functions and the autopersistence graphs of binary autoregressive time series. *Journal of Time Series Analysis*, 32(6):639–646, 2011.

[49] White, H. Connectionist nonparametric regression: Multilayer feedforward netwoks can learn arbitrary mappings. *Neural Networks*, 3:535–549, 1990.

[50] Wilks, D., and Wilby, R. The weather generation game a review of stochastic weather models. *Progress in Physical Geography*, 23:329–357, 1999.