Monitoring time series based on estimating functions

Claudia Kirch* Joseph Tadjuidje Kamgaing†

January 23, 2014

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 even 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, allowing 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

*Karlsruhe Institute of Technology (KIT), Institute for Stochastics, Kaiserstr. 89 D–76133 Karlsruhe, Germany; claudia.kirch@kit.edu

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

1 Introduction

In recent years an increasing number of data sets are collected automatically or without significant costs in such a way that the observations arrive steadily. Examples include financial data sets, e.g., in risk management (Andreou and Ghysels [1]) or CAPM models (Aue et al. [3]) as well as medical data sets, e.g., monitoring intensive care patients (Fried and Imhoff [14]). More applications can be found in different areas of applied statistics. The consideration of such data sets leads to sequential statistical analysis, which is also called online monitoring.

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.

Chu et al. [7] introduced a new way of sequential testing, which allows to control the asymptotic α-error if no changes occur while having asymptotic power one under alternatives, which has then be pursued by others (confer Section 6 below for many examples). This is in contrast to classical sequential change-point procedures which usually solve an optimization problem under restrictions in a fully parametric setup such as minimizing the average delay time of the alarm under restrictions on the lower bound of the average time until alarm if no change occurs. This latter approach has a different ideology behind it and usually leads to tests that will reject eventually even if nothing occurs but with the merit of having a small detection delay.

In the approach of Chu et al. [7] the existence of a historic data set without change is assumed. In applications such a data set almost always exists as at least some data needs to be collected before any reasonable statistical inference can take place. In the context of sequential change-point tests, this data set is also used to estimate the parameter of interest which is the parameter that we want to monitor for changes as well as some additional parameters which influence the asymptotics such as the variance. Asymptotic theory can then be derived even for a possibly infinite monitoring horizon by letting the length of this historic data set grow to infinity. Due to the asymptotic framework it is not necessary to make any additional parametric assumptions for example on the error distribution of an autoregressive time series and one can even deal with misspecification confer Section 6.3 below.

Using this same idea, we will then extend the existing literature by deriving a general construction principle for such sequential change-point tests based on estimating functions, which is sometimes also called generalized method of moments, in Section 2. In Sections 3 and 4 we will then 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 structure as illustrated in Section 5. In Section 6 we will then give relevant examples of time series and estimating functions which fulfill these regularity conditions. This

includes examples already discussed in the literature, for which the class of applicable tests is somewhat extended by the very general framework considered in this paper. Relevant new examples include tests for nonlinear autoregressive time series in addition to tests for certain count time series. The new examples in this section are accompanied by simulations as well as data examples. 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, whose asymptotics will then be derived in the following sections. For readers, who are not already familiar with this methodology it may be helpful to have 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 function class leading to well defined stationary time series and $\{e_t\}$ is either i.i.d. or allows for some additional time series structure. In this situation, i.i.d. errors correspond to the correctly specified case, which is usually considered in the literature. An additional time series structure of $\{e_t\}$ arises for example in the misspecified case, where we merely use the parametric model f_θ to approximate a non-parametric function f in order to construct change-point tests. Such a semi-parametric approach will then only be able to detect changes for which the best approximating parameters before and after the change are different. A similar idea in the a posteriori setup has been considered by Kirch and Tadjuidje Kamgaing [23].

The following derivation is not restricted to the above two examples but it may be helpful to keep them in mind. Before we can start monitoring, we need to have some historic data which is stationary and does not contain a change. This is called the 'non-contamination assumption' by Chu et al. [7]. This assumption is usually fulfilled in applications, since statistical inference, such as prediction amongst other, is typically only carried out after a model has been constructed which describes existing data well enough. In this same spirit, we will use this historic data set to estimate the unknown parameter of interest, before starting to monitor future incoming observations, deciding whether this same parameter still describes the new observations well enough. Because we allow this monitoring to continue for ever (in the open end procedure below) if no alarm is raised, it is important to understand that the asymptotic considerations in the following two section are obtained with respect to the length of the historic data set increasing to infinity, which means that the estimator used in the monitoring becomes increasingly accurate.

A large class of estimation procedures are based on estimating functions sometimes also called objective function in the generalized method of moments framework, where an

estimator is obtained as the solution of the following system of equations:

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

where $\mathbf{X}_t, t = 1, \ldots, m$, are the historic observations and G is a suitable estimating function with values in $\mathbb{R}^d$, where d is the number of unknown parameters in the parametric representation of interest. 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 true parameter θ_0, which fulfills $\mathbb{E}G(\mathbf{X}_1, \theta_0) = 0$ in the correctly specified case and for the best approximating parameter θ_0 in the sense of $\mathbb{E}G(\mathbf{X}_1, \theta_0) = 0$ under misspecification. Standard examples include (weighted) least-squares, ML- or M-estimation.

After having estimated the unknown parameters in the model, we start monitoring new incoming observations for a change in those parameters. To this end we use a monitoring function H, for which $\mathbb{E}H(\mathbf{X}, \theta_0) = 0$ with θ_0 defined by $\mathbb{E}G(\mathbf{X}, \theta_0) = 0$. This monitoring function can be the same or a different estimating function but can also be of lower dimension d'. In the latter case restriction apply which alternatives are detectable by the corresponding sequential tests with the merit of increased power for some alternatives. 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 often, but not always, part of the estimating equations given by G.

The monitoring statistic is based on

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

where $\widehat{\theta}_m$ is the estimator from (2.1) based only on the historic data set $X_1, \ldots, X_m$. 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 showing that, $\boldsymbol{S}(m, k)$ should be small. On the other hand if a change (in the best approximating parameter) 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. We will formalize these rather vague statements by deriving exact asymptotics in Sections 3 and 4. Since $\boldsymbol{S}(m, k)$ is possibly a vector, the corresponding monitoring scheme will be based on a quadratic form, hence we reject as soon as

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

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. 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:

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

where $N(m) = \infty$ in case of an open-end procedure and $N(m) = Nm + 1$, $N > 0$, for the closed-end procedure. If $\tau(m) = \infty$ we did not reject the null hypothesis during the observation period. Otherwise, it tells us at what time the null hypothesis was rejected and the procedure stopped.

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 already mentioned the solution proposed by Chu et al. [7] is to use asymptotics with respect to the length m of the historic data set. 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 this monitoring procedure detects a large class of alternatives with probability 1 asymptotically, i.e.

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

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

The detection power in such procedures will largely be influenced by the choice of estimating and monitoring function. The more precise the estimators are obtained from the estimating functions respectively the clearer the monitoring function can distinguish between different parameter values, the better the detection power of the corresponding procedure will be. In this sense, using maximum likelihood scores will typically be preferably to using least squares scores. On the other hand, different properties such as robustness properties will 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 will then give examples where those regularity conditions are fulfilled.

First, we need to 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,

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

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 of the location of the change point. This is due to the fact that we stop if the partial sum process $\boldsymbol{S}(m,k)\boldsymbol{A}\boldsymbol{S}(m,k)$ crosses the boundary function $cw(m,k)^{-1}$ (for $w(m,k) \neq 0$) for the first time. If a second boundary function is e.g. below the first one in the region after the change point, detection will be quicker. Two boundary function that control the size at a given level in the sense of (2.3) will always cross each other at least once. Consequently, this quicker detection for certain change locations leads to longer detection delay for other change locations.

A. 2. The following approximation holds under H_0, where $N(m)$ is the possibly infinite observation horizon:

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

for some θ_0, where $\boldsymbol{B}(\theta_0)$ depends on the distribution of $\mathbf{X}_1$ as well as θ_0. 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 the same estimating functions are used or a projection onto one particular component of the estimating functions (such as estimated residuals for many but not all estimating functions). Examples will be 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.

A. 3. a) (i) 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 for any $T > 0$ a functional central limit 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

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

(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

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

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 a strong invariance principle, i.e. (possibly after changing the probability space) there exists a Wiener process $\boldsymbol{W}$ with covariance matrix $\boldsymbol{\Sigma}$ as in (3.3) such that

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

c) To obtain the below 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 $\{\mathbf{W}_2(\cdot)\}$ exists, each with covariance matrix Σ_1, such that for some $\nu > 2$ it holds

$$\sup_{k\geqslant 1} \frac{1}{k^{1/\nu}} \left(\sum_{t=m+1}^{m+k} H(\mathbf{X}_t, \theta_0) - \mathbf{W}_2(k) \right) = O_P(1),$$
$$\frac{1}{m^{1/\nu}} \left(\sum_{t=1}^{m} H(\mathbf{X}_t, \theta_0) - \mathbf{W}_1(m) \right) = O_P(1).$$

This reduces to the corresponding assumptions on $H(\mathbf{X}, \theta_0)$ in the standard situation, where $\boldsymbol{B}(\theta_0)G(\mathbf{X}, \theta_0) = H(\mathbf{X}, \theta_0)$. 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. Such an argument typically involves some kind of cutting or big-block-small-block argument. Details on how to prove it for augmented GARCH time series can be found in Aue et al. [2].

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

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

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) yield 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 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 $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ 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

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

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 $B(\theta)G = H$ *with 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*

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

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

$\Gamma(\cdot)$ *is the Gamma-function and* d' *the dimension of the estimating 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).*

If $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$, e.g. for $\boldsymbol{B}(\theta_0)G(\mathbf{X},\theta_0) = H(\mathbf{X},\theta_0)$, then often $\mathbf{A} = \mathbf{\Sigma}_1^{-1}$ is chosen leading to a pivotal limit.

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* $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$*, then for any* $0 \leqslant \gamma < 1/2$

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

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

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

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

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

4 Consistency under alternatives

In this section, we state some regularity conditions under which changes are asymptotically detected with power one meaning that the procedure will eventually stop if a change does occurs. As discussed below Condition A.1 the detection delay time is influenced mainly by the weight function that is used in combination to the location of the change.

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 exists a ball $U(x_0)$ around x_0 with $x_0 > \vartheta$ and $\rho(x) \geqslant c > 0$ for $x \in U(x_0)$ as well as

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

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 $l \to \infty$ it holds

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

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, then this holds true if the time series before and after the change have different best approximating parameters (in the sense of G and H respectively). However, if only a subset of estimating equations is used in the detection procedure, this is a restriction which can lead to increased power for particular alternatives. 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) and b) and* $\mathbf{A}^{1/2}\mathbf{E}_H \neq 0$*, the closed-end procedure has asymptotic power one, hence will eventually stop. The open-end procedure has asymptotic power one under A.4 a) and either b) or c) as well as* $\mathbf{A}^{1/2}\mathbf{E}_H \neq 0$*. 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

5.1 Conditions under the null hypothesis

If the estimating and monitoring functions are sufficiently smooth and under some mild regularity conditions on the underlying null time series, we get Assumption A.2 above. Some robust estimating functions of interest (such as L_1-minimizer) are not smooth

so a different approach is needed in order to obtain A.2 (confer Section 6.2 below). Alternatively, they can often be approximated to any degree of accuracy by estimating functions fulfilling the smoothness conditions stated here.

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 $\|\theta-\theta_1\|>\epsilon$.

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

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

and $\mathbb{E}\nabla G(\mathbf{X}_1,\theta_0)$ is positive definite.

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

The first assertions are regularity conditions, 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 get 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

Assume that at point $m + k^*$ a change occurs. In many situations such as regression situations, the following conditions are fulfilled:

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

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

Consequently, 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 we get* (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 of 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 $\|H(\mathbf{X}_j, \widehat{\theta}_m) - H(\mathbf{X}_j^*, \widehat{\theta}_m)\| \leqslant \|\boldsymbol{R}_t F(\boldsymbol{X}_t^*)\| + C\|\boldsymbol{R}_t\|^2$ for some measurable function F such that $\mathbb{E}\|F(\boldsymbol{X}_t^*)\|^2 < \infty$.

Assumption a) allows for example 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. [18]. For linear autoregressive models this is naturally fulfilled and can easily be checked, for non-linear autoregressive settings, some work need 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 [32] in addition to condition (iv) in Theorem 21.12 in Lindvall [31]).

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 will both give a survey of existing results where the above kind of monitoring scheme has been proposed and the regularity conditions proven as well as give new examples, that have not yet been discussed in the literature. Because of the great generality of the considered weight functions in this paper, we automatically 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 some simulations and some data examples are also included. The empirical results of the simulations are always based on 1000 repetitions.

6.1 Linear regression

Consider the classical linear regression model

$$Y_i = \mathbb{Z}_i^T \boldsymbol{\beta}_0 + \varepsilon_t,$$

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

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{Z}_i \mathbb{Z}_i^T - \boldsymbol{C} = O(n^{-\tau}) \qquad a.s. \tag{6.1}$$

The papers of Chu et al. [7] and Horváth et al. [17] 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.

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

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

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

Since by assumption $Z_{t,1} = 1$, the monitoring function H is the first line of G, hence we get $\boldsymbol{B}(\boldsymbol{\beta}_0)G = H$, Horváth et al. [17] then 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 $H((Y_t, \mathbb{Z}_t^T), \boldsymbol{\beta}_0) = \varepsilon_t$ 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. [27, 28]. Extensions to the non-i.i.d. case have been proposed by Aue et al. [4] for certain martingale difference sequences including augmented GARCH processes as well as by Schmitz and Steinebach [35] for certain weak dependent processes. Horváth et al.[19] 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. [19] shows (4.2) with $\boldsymbol{E}_H = \boldsymbol{c}_1^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)$ if $Y_t = \mathbb{Z}_t^T\boldsymbol{\beta}_1 + \varepsilon_t$ after the change, where $\boldsymbol{c}_1$ is the first column of the matrix $\boldsymbol{C}$ as in (6.1). Consequently, only changes are detectable for which $\boldsymbol{E}_H \neq 0$ which essentially means that the change goes along with a mean change.

Because of the above power restriction Hušková and Koubková [21] proposed to use the monitoring function

$$H((Y_t, \mathbb{Z}_t^T), \boldsymbol{\beta}) = \mathbb{Z}_t(Y_t - \boldsymbol{\beta}^T\mathbb{Z}_t) \tag{6.2}$$

proving the above assumptions. For this choice, all alternatives have asymptotic power one because $\boldsymbol{E}_H = (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^T\mathbb{E}\mathbb{Z}_{k^*+1}$, which is under alternatives always different from zero.

6.2 Mean changes

Obviously, monitoring schemes for mean changes are already included in the linear regression setup from the section before with $\mathbb{Z}_t = 1$, leading to monitoring procedures based on the sample mean.

Because that statistic is not particularly robust, extensions to more robust M-estimators have been considered in the dissertations of Koubková [29] as well as Chochola [6]. Koubkova [29] (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 then be obtained by standard methods on the i.i.d. errors $\operatorname{sgn}(X_i - \mu_0)$ respectively $\psi(X_i - \mu_0)$. Koubková [29] also extends her results to the linear regression situation of the previous subsection. Chochola [6] also 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 [29] and Chochola [6] consider general M-estimators, which are not necessarily differentiable (or even continuous) as the important example of the median with the estimating function $\operatorname{sgn}(X_i - \mu)$ shows, the methodology provided by Proposition 5.2 cannot be applied and a different approach is necessary to obtain A.2. However, for M-estimators with sufficiently smooth ψ-functions, the methodology provided by Proposition 5.2 can easily be used to provide the asymptotic theory for corresponding monitoring schemes.

In particular, the theory derived in this paper allows to use less robust but more precise estimators such as the sample mean to get a parameter estimator based on the historic data set, while using a more robust monitoring function. Such a procedure can be important in practice if 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. If the regularity assumptions of Proposition 5.2 are not fulfilled (as e.g. for the sign-function and the median), some additional work is needed 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):

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

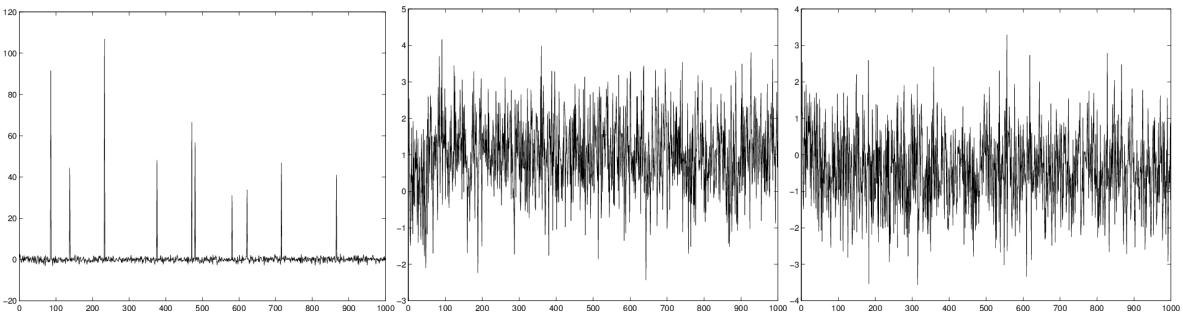
By $\frac{\partial \tanh(u)}{\partial u} = 1 - \tanh^2(u)$ and Proposition 5.2 it follows $\mathbf{B}(\mu_0) = \mathbb{E}\tanh^2(X - \mu_0) - 1$, so that we get by Theorem 3.1 b) and Theorem 3.2 b)

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

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 changes in the mean. In Figure 6.1 a sample path for each of the latter three time series is given.



(a) Data with outliers (b) $\mu_1 = 1$ (c) $\mu_1 = -0.5$

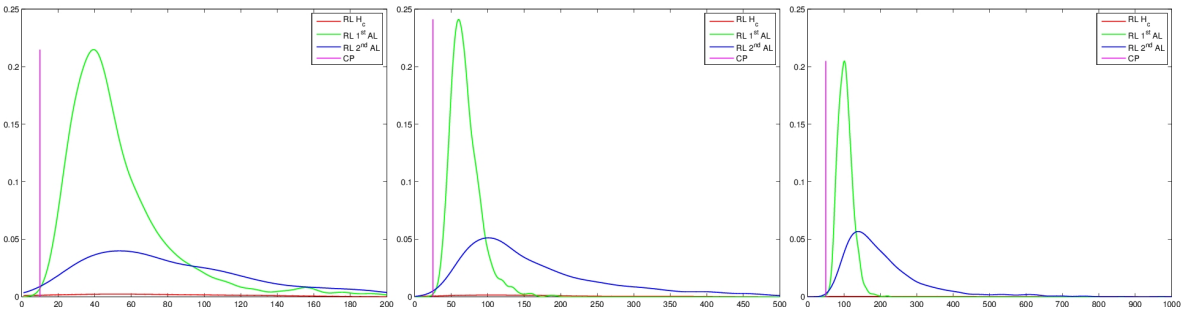
Figure 6.1: Sample Monitoring data: m=100

In the simulations, we stop the 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 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. In this plot we also include the results for the null hypothesis including outliers (H_c) to indicate that the test is really robust with respect to outliers.

Table 6.1: Empirical size and power: Mean change

	$H_0 : \mu_1 = 0$			H_c			$H_1 : \mu_2 = 1$			$H_1 : \mu_2 = -0.5$		
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



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

Figure 6.2: Scaled density estimate of the run length

The procedure is conservative with the 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. [5] use the log likelihood score function as estimating as well as monitoring function. Lemma 6.4 in that work proves Assumption 2, the proof of Lemma 6.6 shows A.3 (i) – (iii).

Ciuperca [8] considers a nonlinear regression model $Y_i = f(\mathbb{Z}_i, \boldsymbol{\beta}_0) + \varepsilon_i$ 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

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

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

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

Since this is not an estimating function, we 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 essentially 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 closed-end procedure with the standard weight function in (3.2). Unlike in the linear regression situation H is no longer necessarily a linear combination of the components of G resulting in a situation, where $\boldsymbol{B}(\theta_0)G \neq H$. In this situation the weight function (3.2) does no longer lead to a pivotal limit with all the problems this entails. However, a weight function as is implicit in Theorem 3.2 can solve that problem leading to a pivotal limit. Since the regularity conditions of Ciuperca [8] are similar to the ones given in Section 5, $\boldsymbol{B}(\theta_0)$ is given 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. [27, 28].

Kirch and Tadjuidje Kamgaing [24] use a similar monitoring scheme for autoregressive time series, where the nonlinear function g is given by a neural network taking possible misspecification into account. This is the equivalent of the offline procedure considered in Kirch and Tadjuidje Kamgaing [23]. Their idea is to construct a test that is able to detect a large class of alternatives in general non-linear nonparametric autoregressive time series. Since neural network functions can approximate a large class of functions to any degree of accuracy (confer e.g. White [40] or Franke et al. [13] and some of the references therein,) this amounts to choosing a parametric approximation of a nonparametric regression function to be used in the tests. Since we cannot and do not expect the time series to actually follow that precise model, we need to take misspecification into account. This is very similar to the idea of using sieve estimators in nonparametric statistics. Consequently, changes can be detected that result in different best approximating parameters for the time series before and after the change with possible restrictions if the monitoring function is no estimating function.

To elaborate, we assume that the data are realization of an autoregressive process Y_t with

$$Y_t = g(Y_{t-1}, \ldots, Y_{t-p}) + \varepsilon_t, \tag{6.3}$$

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

Under these assumptions Condition B.1 is fulfilled. For estimating and monitoring functions as given below, Conditions B.2 and B.3 are also fulfilled, which by Proposition 5.2 yields A.2. Since for the below estimating and monitoring functions G and H it holds $\boldsymbol{B}(\theta_0)G = H$, Condition A.3 b) follows from the invariance principle of Kuelbs and Philipp [30] 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. Nevertheless, for the nonlinear autoregressive model, assuming for example that the ϵ_t are realizations from a random variable with density that is strictly positive on the real line, one could

make use of the standard Markov chain stability theory (see e.g. Meyn and Tweedie [32] or Tong [38]) to derive the key property of geometric ergodicity for the process X_t. 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 then based on ϕ-irreducibility, aperiodicity and the drift condition for Markov chain, that need to be guaranteed (see Chapter 15 of Meyn and Tweedie [32]). Such application can be found in Stockis et al. [37], in a broader framework, they use autoregressive in time series based on neural network functions as building blocks in a regime-switching model, so called CHARME-models, in the context of financial time series. However, We are interested in monitoring for changes in the autoregression function g. In order to construct a monitoring scheme, we use the estimating functions for a parametric approximation based on a one layer feedforward neural network with n_H hidden neurons

$$f(\mathbf{x}, \theta) = \nu_0 + \sum_{h=1}^{n_H} \nu_h \psi(< \boldsymbol{\alpha}_h, \mathbf{x} > + \beta_h), \tag{6.4}$$

where $\theta = (\nu_0, \ldots, \nu_{n_H}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{n_H}, \beta_1, \ldots, \beta_{n_H}) \in \Theta$ and we assume Θ to be convex and compact, $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jp})$ 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

$$\lim_{x \to -\infty} \psi(x) = 0, \qquad \lim_{x \to \infty} \psi(x) = 1, \qquad \psi(x) + \psi(-x) = 1. \tag{6.5}$$

A popular example is the logistic function $\psi(x) = (1 + e^{-x})^{-1}$. Since neural network functions can approximate a large class of functions to any degree of accuracy, see, e.g., White [40] or Franke et al. [13]. To this end, define the estimating function (for ordinary least squares estimation) as

$$G((Y_t, \mathbb{Y}_{t-1}), \theta) = (Y_t - f(\mathbb{Y}_{t-1}, \theta)) \, \nabla f(\mathbb{Y}_{t-1}, \theta), \qquad \mathbb{Y}_{t-1} = (Y_{t-1}, \ldots, Y_{t-p})^T.$$

The dimension of the estimating function G is then given by $d = n_H(p+2)+1$. Since f is twice continuous differentiable by assumption, G is at least 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 $\{Y_t\}$ 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 assumption. It remains to assume B.2 b), which is essentially an identifiability condition, saying that the best approximating parameter θ_0 given by $\mathbb{E}G((Y_t, \mathbb{Y}_{t-1}, \theta_0) = 0$ is identifiable unique. The positive definiteness condition in c) is another regularity condition which is standard in the literature, see for example Hall [15].

By Proposition 5.1 we get $\sqrt{m}$-consistency of the least squares estimator to the best approximating parameter. Kirch and Tadjuidje Kamgaing [24] consider monitoring schemes based on the estimated residuals, i.e. they use as monitoring function

$$H((Y_t, \mathbb{Y}_{t-1}), \theta) = Y_t - f(\mathbb{Y}_{t-1}, \theta). \tag{6.6}$$

Hence $\boldsymbol{B}(\theta_0)G = H$, which fulfills the assumptions in B.3 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.

Alternatively, we can use the full estimating function in the monitoring procedure, i.e. $\tilde{H} = G$ (and $\boldsymbol{B}(\theta_0) = \mathrm{Id}$). In this case, B.3 follows if the activation function ψ is additionally three times continuous differentiable with bounded third derivative due to the existence of fourth moments.

To summarize, 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 test statistics are at hand. If one chooses $\boldsymbol{A}$ as the inverse of the long-run covariance matrix of $H((Y_t, \mathbb{Y}_{t-1}), \theta_0)$ respectively $\tilde{H}((Y_t, \mathbb{Y}_{t-1}), \theta_0)$ one get 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

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

in case of $\tilde{H}$. Because one will usually only apply the procedure if the fit by the neural network is relatively good on the one hand and because estimators for the long-run variance are not very precise on the other hand, it is often better to estimate the above covariance matrix rather than the long-run covariance matrix. Essentially, this trades a large estimation error for a small model error. This is confirmed by the simulations in Kirch and Tadjuidje Kamgaing [23]. Consequently, we propose to use

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

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

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

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

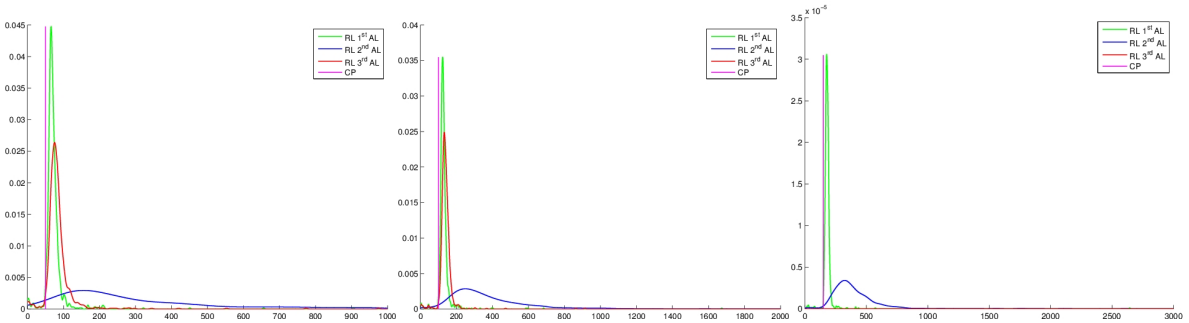
This implies B.5 b) for both H (with $F = D$) as well as $\tilde{H}$ (with $F = D\,\mathrm{id}$) if $\{Y_t^*\}$ has second moments. In this case $\boldsymbol{E}_H = \mathbb{E}(X_t^*) - \mathbb{E} f(\mathbb{X}_t^*, \theta_0)$ showing that essentially mean changes will be detected. On the other hand $\boldsymbol{E}_{\tilde{H}} = \mathbb{E} G(X_t^*, \theta_0)$ which will be 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 a different best approximating parameter in the neural network approximation.

Simulation study

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

	$H_0 : \theta_1 = (1, 0.3)$			$H_1 : \theta_2 = (3, 0.75)$			$H_1 : \theta_2 = (1, 0.75)$			$H_1 : \theta_2 = (3, 0.3)$		
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

In order to illustrate the behavior of the above statistic (based on the neural network approximation as in (6.4) with 2 hidden neurons) under misspecification, we apply it to a first order linear autoregressive process $Y_t = \omega + \alpha Y_{t-1} + \varepsilon_t$, $\{\varepsilon_t\}$ i.i.d. standard Gaussian. We use the monitoring function as given in (6.6) with $\widehat{\mathbf{A}}_H$. Furthermore, we use the extreme-value statistic as in Theorem 3.1 c), where we truncate the simulations at $10m$. 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 quickly. This effect is probably due to the misspecification where a somewhat longer training sample is needed in order for the obtained approximation to be good enough to generalize to future incoming observations. Nevertheless, keeping in mind that the misspecified approximation obtained from the historical sample is used on new data, the obtained level (which corresponds to how good the approximation is) is not bad even for smaller historic samples. The power is good but of course depends on the kind of parameter change present. This becomes even clearer when looking at the scaled density estimate of the rung length as given in Figure 6.3.



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

Figure 6.3: Scaled density estimate of the run length for the misspecified neural network monitoring with a true underlying AR(1)-model

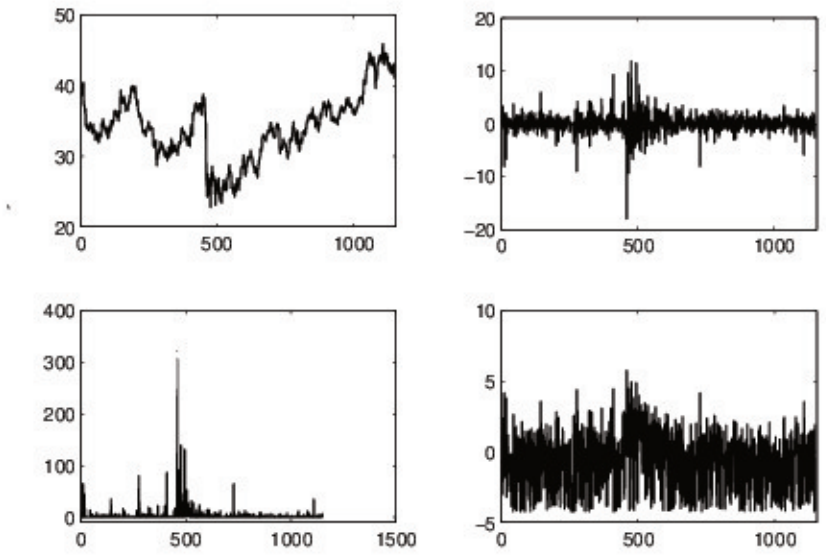


Figure 6.4: SAP stock returns and suitable transformations (January 2007 to April 2011)

Data Example

After a suitable data transformation β-ARCH-models can be treated within this framework, so that we can apply the above methodology for suitably transformed log-returns of financial assets. More details on this can be found in Kirch and Tadjuidje Kamgaing [23]. Here, we apply the methodology to the S&P500 index from January 2007 to April 2011. Figure 6.4 shows the stock price, the log returns, squared log-returns and finally suitably transformed log returns. Using the offline procedure of Kirch and Tadjuidje Kamgaing [23] in combination with a binary segmentation step yields two possible break points at 2.9.2008 as well as 15.05.2009 probably associated with the financial crises. In Figure 6.5 the 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 also raises an alarm quickly after the first possible change point agreeing with the offline procedure.

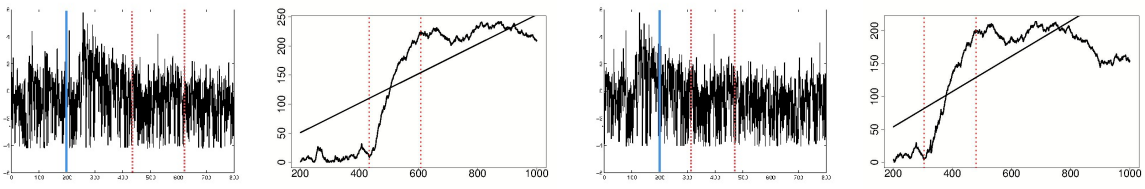


Figure 6.5: Detector for SAP data

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 [41] for example observe, whether it has been raining on a specific day, Kauppi and Saikkonen [22] and Startz [36] observe whether or not a recession has occurred in a given month. A common binary time series model is given by

$$Y_t \mid Y_{t-1}, Y_{t-2}, \ldots, Z_{t-1}, Z_{t-2}, \ldots \sim \text{Bern}(\pi_t(\boldsymbol{\beta})), \qquad \text{with } g(\pi_t(\boldsymbol{\beta})) = \boldsymbol{\beta}^T \mathbb{Z}_{t-1}, \tag{6.7}$$

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

$$G((Y_t, \mathbb{Z}_{t-1}), \boldsymbol{\beta}) = \mathbb{Z}_{t-1}(Y_t - \pi_t(\boldsymbol{\beta})) \tag{6.8}$$

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 $(Y_t, \mathbb{Z}_{t-1})$ is strong mixing with exponential rates. For $\mathbb{Z}_{t-1} = (Y_{t-1}, \ldots, Y_{t-p})^T, Y_t$ is the standard binary autoregressive model (BAR(p)), Wang and Li [39] showed the geometric ergodicity property which in turn implies strong mixing with exponential rates. However, considering some regularity assumptions on the exogenous process, one can prove that $(Y_t, \ldots, Y_{t-p+1}, Z_t, \ldots, Z_{t-q})$ is a Feller chain, for which Theorem 1 of Feigin and Tweedie [10] can be applied to derive its geometric ergodic property (see Kirch and Tadjuidje Kamgaing [25] for details on this issue). Alternatively, invariance principles based on results of Eberlein [9] can be used (for details we refer to Fokianos et al. [11], 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 Y_t^* and $\pi_t(\boldsymbol{\beta})$, while the same arguments as in the proof of Theorem 1.3.2 in Kirch and Tajduidje Kamgaing [26] give B.5 b). If both Y_t and Y_t^* follow BAR-models, all parameter changes are thus detected by the identifiability of the parameter via the partial score function.

Simulation study

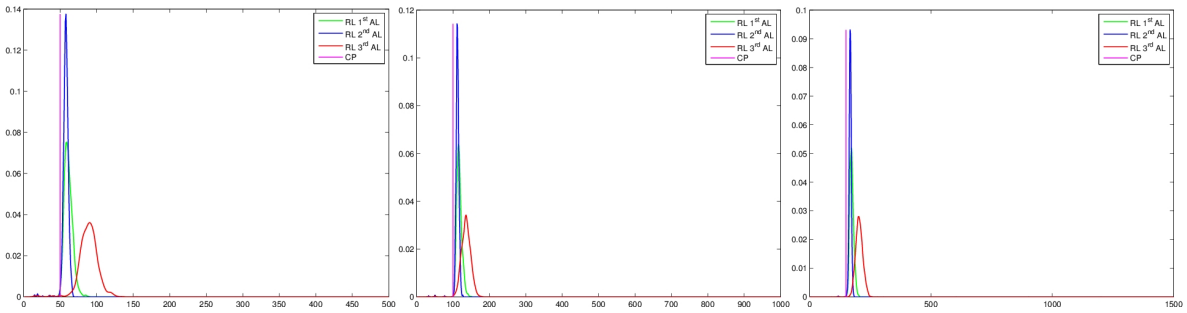
We will now illustrate the monitoring for binary data by using a first order binary autoregressive process as in (6.7) with $\mathbb{Z}_{t-1} = (1, Y_{t-1})$. We use the closed-end monitoring statistic with $G = H$ as above with $N = 5$ and $w(m,k) = m^{-1/2}\left(1 + \frac{k}{m}\right)^{-1}$. We can 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 $\widehat{\beta}$ is estimated based on the estimation function G and the historic data set only. Theorem 3.1 gives the null asymptotics in this situation.

	$H_0 : \beta_1 = (2, -2)$			$H_1 : \beta_2 = (-2, 2)$			$H_1 : \beta_2 = (-3, -2)$			$H_1 : \beta_2 = (2, 1)$		
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



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

Figure 6.6: Scaled density estimate of the run length for BAR(1)-model

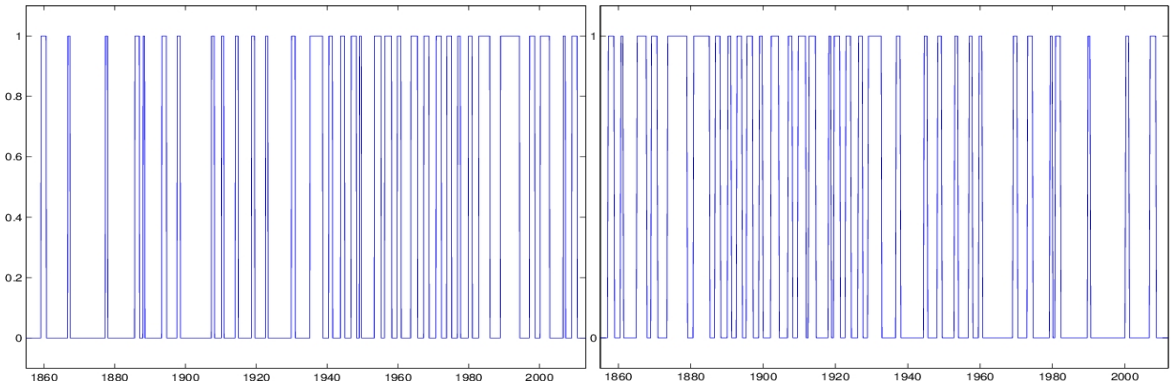
Table 6.4 reports the empirical size and power (based on 1000 repetitions) for the nominal 5% level and various alternatives, where a change always occurred at time $\frac{m}{2}$ after the monitoring started. Figure 6.6 gives the scaled density estimator of the run length. The monitoring is conservative under the null hypothesis and detects the considered changes with empirical power one and relatively quickly after they occur.

Data Example

We now 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] The quarterly version of this data set has been analyzed by Kirch and Tadjuidje Kamgaing [26] and Hudecová [20] in the context of offline change point detection. Their findings indicate the existence of a change point in the 1930s, where we obtained similar findings on the monthly data.

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 [26]). The corresponding detectors can be found in Figure 6.8. Three out of four of the detector schemes detect the change point until now, where the detection delay is shorter the later we start monitoring, which is mainly due to the fact that the change point occurs relatively late after monitoring begins, which seems to be particularly problematic in this

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



(a) Monthly Data (b) Quarterly Data

Figure 6.7: US recession data

situation of binary data. This can be alleviated by using somewhat different detector sums with a shorter memory which will be investigated in future work. For the monitoring of quarterly data with $m = 200$ a clear increase in the detector is visible but it is has not yet rejected.

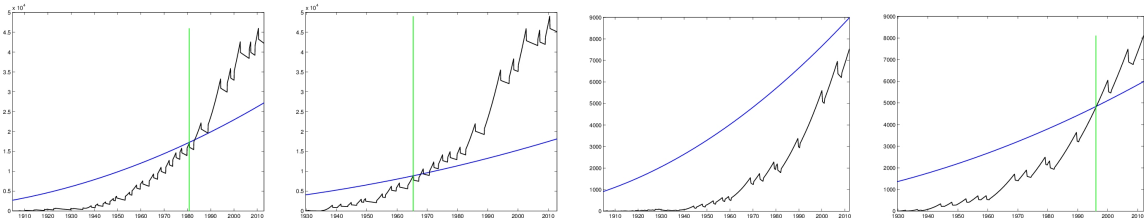
6.5 Poisson autoregressive time series

Another popular model for time series of counts is given by the Poisson autoregression, where we observe $Y_1, \ldots, Y_n$ with

$$Y_t \mid Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p} \sim \text{Pois}(\lambda_t), \qquad \lambda_t = f_\theta(\mathbb{Y}_{t-1}), \qquad \mathbb{Y}_{t-1} = (Y_{t-1}, \ldots, Y_{t-p})^T. \tag{6.9}$$

If $f_\theta(\mathbf{x})$ is Lipschitz-continuous in $\mathbf{x}$ for all $\gamma \in \Theta$ with Lipschitz constant strictly smaller than 1, then there exists a stationary ergodic solution of the (6.9) which is β-mixing with exponential rate (confer Neumann [33]). 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



(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

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. [12] 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.

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

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 if the second moment of the process X_t exists, implying B.2a).

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 if third moments exist. Similarly,

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

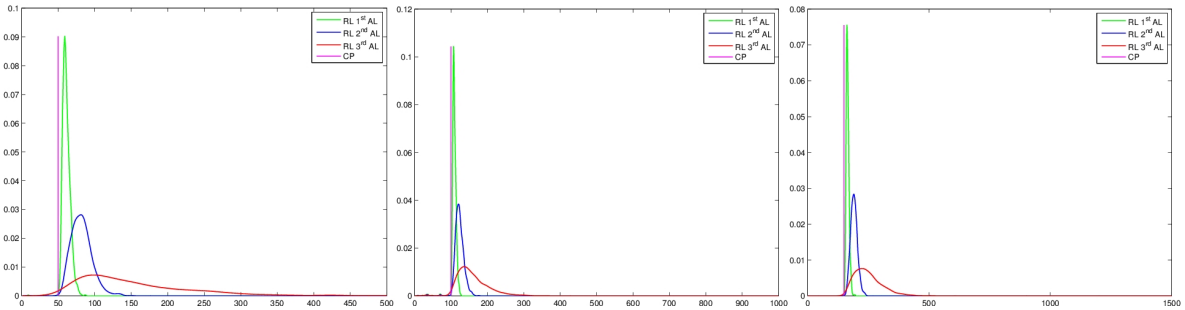
if fourth moments exist, yielding B.3 for $H = G$.

	$H_0 : \theta_1 = (1, 0.5)$			$H_1 : \theta_2 = (3, 0.75)$			$H_1 : \theta_2 = (3, 0.5)$			$H_1 : \theta_2 = (1, 0.75)$		
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)

Simulation study

In the simulations we consider a Poisson autoregressive model as in (6.9) with $\lambda_t = \theta_1 + \theta_2 Y_{t-1}$. 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 $\boldsymbol{S}_1 = \boldsymbol{S}_2 = \mathbb{E}\mathbb{Z}_{t-1}\mathbb{Z}_{t-1}^T \frac{(Y_t - \lambda_t)^2}{\lambda_t^2}$ by the empirical covariance matrix of $\{G((X_t, X_{t-1}), \hat{\theta})\}$. Theorem 3.1 gives the null asymptotics in this situation.



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

Figure 6.9: Scaled density estimate of the run length for BAR(1)-model

Table 6.5 shows the empirical size and power at the nominal 5% level for various alternatives, while Figure 6.9 shows the scaled density. The tests are conservative and reject the null hypothesis for all but one simulated alternatives, however, the detection delay differs for different alternatives with the smallest change having the longest detection delay.

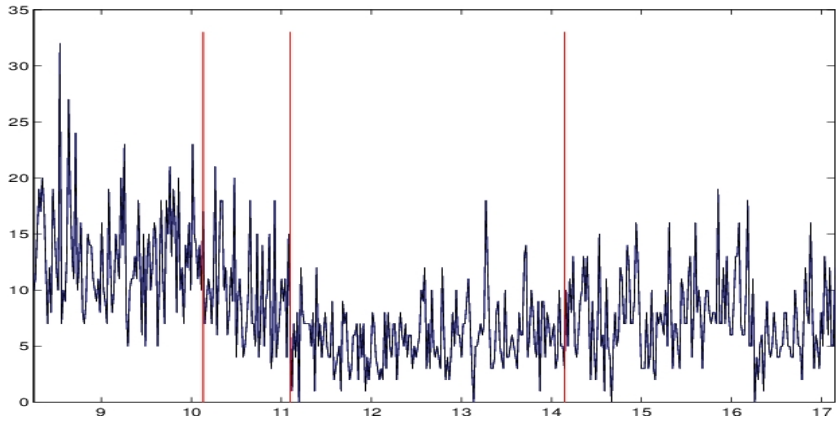
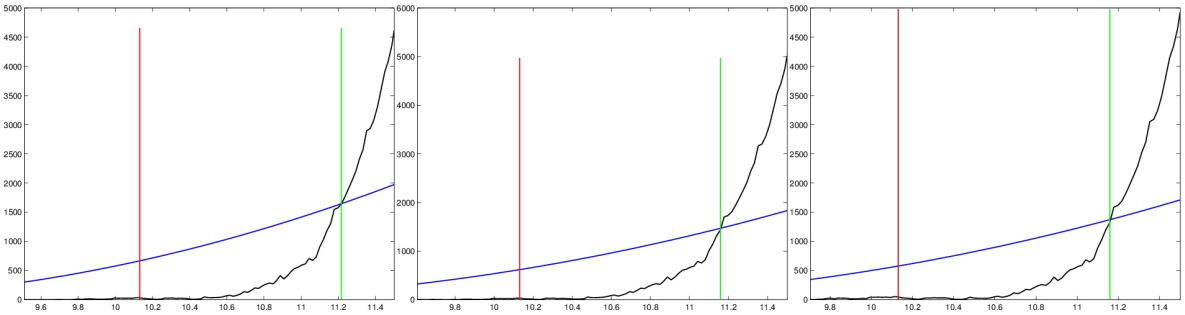


Figure 6.10: Stock Ericsson B: July 3rd 2002 and possible change points (by binary segmentation)

Data 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

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. Kirch and Tadjuidje [26] 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 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).

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. Introduce the notation $\|Z\|_{\boldsymbol{A}}^2 = |Z^T \mathbf{A} Z|$, 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 $N > 0$

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

Noting that $\{\mathbf{W}_1(1+t) - \mathbf{W}_1(t) : t \geqslant 0\}$ is again a Wiener process with covariance matrix $\boldsymbol{\Sigma}_1$ 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 the $\sup_{\tau\leqslant t\leqslant N} \rho^2(t)\,\|\mathbf{W}_1(1+t) - \mathbf{W}_1(1) - t\mathbf{W}_2(1)\|_{\boldsymbol{A}}^2$ 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 $\gamma < \alpha < 1/2$

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

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

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

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

$$\begin{aligned} &\sup_{k \geqslant \tau m} \left\| w(m, \min(k, Tm)) \left(\sum_{j=m+1}^{m+\min(k,mT)} H(\mathbf{X}_j, \theta_0) - w(m,k) \frac{k}{m} \boldsymbol{B}(\theta_0) \sum_{j=1}^{m} G(\mathbf{X}_j, \theta_0) \right) \right\|_{\boldsymbol{A}}^2 \\ &\xrightarrow{\mathcal{D}} \sup_{t \geqslant \tau} \left\| \rho(\min(t,T))(\mathbf{W}_1(1+t) - \mathbf{W}_1(1)) - t\rho(t)\mathbf{W}_2(1) \right\|_{\boldsymbol{A}}. \end{aligned} \tag{7.3}$$

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

The proof of c) is analogous to Horváth et al. [18], 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

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

By the proof of Lemma 2.2 in Horváth [16] 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

$$\begin{aligned} &a(\log m)|a(\log m) - a(\log((m + a(m))/a(m)))| \to 0, \\ &b_d(\log m) - b_d(\log((m + a(m))/a(m))) \to 0, \end{aligned}$$

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

Proof of Theorem 3.2. Part a) can be found in Hušková and Koubkova [21], proof of Theorem 2.1, confer also Horváth et al. [17] for the univariate case. Part b) proceeds analogously with the only difference being that for the Wiener processes here, we have

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

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

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

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

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 [26], the proof of b) is analogous to Theorem 3 in Kirch and Tadjuidje Kamgaing [23]. ■

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

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

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

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

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 [34], 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

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

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 $0 \leqslant \gamma < 1/2$

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

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 follows 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 [31] 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 position of the first author 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] Andreou, E., and Ghysels, E. Monitoring disruptions in financial markets. *Journal of Econometrics*, 135:77–124, 2006.

[2] Aue, A., Berkes, I., and Horváth, L. Strong approximation for the sums of squares of augmented garch sequences. *Bernoulli*, 12:583–608, 2006.

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

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

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

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

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

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

[9] Eberlein, E. On strong invariance principles under dependence assumptions. *The Annals of Probability*, 14:260–270, 1986.

[10] 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.

[11] 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.

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

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

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

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

[16] 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.

[17] Horváth, L., Hušková, M., Kokoszka, P., and Steinebach, J. Monitoring changes in linear models. *Journal of Statistical Planning and Inference*, 126:225–251, 2004.

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

[19] Horváth, L., Kokoszka, P., and Steinebach, J. On sequential detection of parameter changes in linear regression. *Statistics and Probability Letters*, 77(9):885–895, 2007.

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

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

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

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

[24] Kirch, C. and Tadjuidje Kamgaing, J. An online approach to detecting changes in nonlinear autoregressive models. *Discussion Paper, University of Kaiserslautern*, 2011. urn:nbn:de:hbz:386-kluedo-27725.

[25] Kirch, C. and Tadjuidje Kamgaing, J. Geometric ergodicity of binary autoregressive models with exogenous variables. *Discussion Paper, University of Kaiserslautern*, 2013. urn:nbn:de:hbz:386-kluedo-36475.

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

[27] Komlós, J., Major, P., and Tusnády, G. An approximation of partial sums of independent rvs and the sample df. i. *Probability theory and related fields*, 32:111–131, 1975.

[28] Komlós, J., Major, P., and Tusnády, G. An approximation of partial sums of independent rvs and the sample df. ii. *Probability theory and related fields*, 34:33–58, 1976.

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

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

[31] 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.

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

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

[34] Ranga Rao, R. Relation between weak and uniform convergence of measures with applications. *The Annals of Mathematical Statistics*, 33:659–680, 1962.

[35] 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.

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

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

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

[39] 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.

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

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