An online approach to detecting changes in nonlinear autoregressive models*

Claudia Kirch† Joseph Tadjuidje Kamgaing‡

October 11, 2011

Abstract

In this paper we develop monitoring schemes for detecting structural changes in nonlinear autoregressive models. We approximate the regression function by a single layer feedforward neural network. We show that CUSUM-type tests based on cumulative sums of estimated residuals, that have been intensively studied for linear regression in both an offline as well as online setting, can be extended to this model. The proposed monitoring schemes reject (asymptotically) the null hypothesis only with a given probability but will detect a large class of alternatives with probability one. In order to construct these sequential size α tests the limit distribution under the null hypothesis is obtained.

Keywords: Change analysis, nonparametric regression, neural network, autoregressive process, sequential test

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

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

*The work was supported by the DFG graduate college 'Mathematics and Practice' as well as by the DFG grant KI 1443/2-1. The position of the first author was financed by the Stifterverband für die Deutsche Wissenschaft by funds of the Claussen-Simon-trust.

†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

financial data sets e.g. in risk management (Andreou and Ghysels [1]) or CAPM models (Aue et al. [2]) as well as medical data sets e.g. monitoring intensive care patients (Fried and Imhoff [5]). 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 appropriate anymore or the condition of the patient in intensive medical care might have changed.

In this paper we focus on nonlinear autoregressive time series, where we model the autoregression function by a neural network. Due to its universal approximation property, a large class of functions can be approximated by a neural network to any degree of accuracy (confer e.g. White [18] or Franke et al. [4] and some of the references therein). Therefore, this setup is very general and able to model many real-life time series while – at the same time – being mathematical feasible and computationally easier to handle due to its parametric nature.

Stockis et al. [17] use these time series as building blocks in a regime-switching model, so called CHARME-models, in the context of financial time series. In their model the duration time in each regime is random and given by a hidden Markov model, while in classical change-point analysis the duration time is usually fixed and deterministic.

Motivated by these time series Kirch and Tadjuidje-Kamgaing [11] developed offline or a-posteriori change-point tests in such a setup for at-most-one deterministic change-point.

In the spirit of CHARME-models, where a new change occurs from time to time, it is of great importance to use monitoring schemes in order to be able to react to such a change as fast as possible after it occurs. This is also important in view of financial data sets, which have recently been argued to be well modeled by $\sigma_t X_t$ with piecewise constant volatility σ_t and an autoregressive error process X_t. After a log-transformation this model fits nicely into our context.

In this paper we address the development of sequential change-point tests where the in-control process follows is can be approximated by an autoregressive process with a neural network as regression function. Asymptotics are derived under correct specification, where the regression function is indeed given by a neural network as well as under misspecification, where the time series has a different structure.

Our monitoring schemes are related to sequential tests first introduced by Chu et al. [3] and further investigated by Horváth et al. [6] both to detect changes in linear regression models.

For $\theta = (\nu_0, \ldots, \nu_H, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_H, \beta_1, \ldots, \beta_H)$, $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jp})$,

$$f(\mathbf{x}, \theta) = \nu_0 + \sum_{h=1}^{H} \nu_h \psi(< \boldsymbol{\alpha}_h, \mathbf{x} > + \beta_h), \tag{1.1}$$

denotes a one layer feedforward neural network with H hidden neurons, $<,>$ is the classical scalar product on $\mathbb{R}^p$. In this paper we assume that ψ is twice continuously differentiable and belongs to the class of sigmoid activation functions that satisfy

$$\lim_{x\to-\infty}\psi(x)=0,\qquad \lim_{x\to\infty}\psi(x)=1,\qquad \psi(x)+\psi(-x)=1. \tag{1.2}$$

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

The autoregressive time series model with neural network regression function is then given by

$$Z_t = f\left(\mathbb{Z}_{t-1},\theta_0\right)+\varepsilon_t, \tag{1.3}$$

where $\mathbb{Z}_{t-1}=(Z_{t-1},\ldots,Z_{t-p})$, θ_0 is fixed but unknown, ε_t independent of $\mathcal{F}_{t-1}=\sigma\{Z_u,u\leq t-1\}$ the σ-algebra generated by the observations up to time $t-1$. Furthermore, $\{\varepsilon_t : 1\leqslant t\leqslant n\}$ are centered independent identically distributed random residuals having a positive variance. For the theory below we consider general stationary and ergodic time series $\{Z_t\}$ fulfilling certain assumptions such as the existence of a unique parameter $\widetilde{\theta}_0$ such that

$$\widetilde{\theta}_0=\arg\min_{\theta\in\Theta}\mathbb{E}(Z_1-f(\mathbb{Z}_0,\theta))^2. \tag{1.4}$$

In the correctly specified case one has under weak assumptions that $\widetilde{\theta}_0=\theta_0$. The theory shows that the tests also have correct size under misspecification if one uses a different variance estimator. Furthermore, the tests have asymptotic power one under certain conditions on the type of alternatives. Naturally, the condition depends on the approximation of X_t by the parametric model (1.3).

The setup of the monitoring procedure is as follows. We assume we have a historic data set of length m where no change occurred, i.e.

$$X_t=Z_t,\qquad 1\leqslant t\leqslant m. \tag{1.5}$$

This is called the 'non-contamination assumption' by Chu et al. [3]. Practically this corresponds to the data set (with no change) based on which we decide for a model, which in turn is used for statistical inference such as prediction. In financial applications this is the data set based on which we estimate model parameters such as the volatility. If later a model change occurs this statistical inference is meaningless. Therefore, it is very important to detect changes as soon as it becomes clear the model parameters do not represent the present data anymore.

Hence, we are interested in monitoring the future incoming observations for a change in the conditional mean, i.e. we want to test the null hypothesis

$$H_0: X_t=Z_t\qquad t>m,$$

against the alternative

H_1 : There exists $k^*\geqslant 0$ such that

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

where the distribution of $\{Y_t\}$ is different from that of $\{X_t\}$.

The time k^* where the change occurs is unknown and may depend on m and is called the change-point.

We consider a sequential monitoring scheme based on cumulative sums of estimated residuals similarly as discussed by Chu et al. [3] and Horváth et al. [6] for the linear regression case. Let

$$\Gamma(m,k) = \sum_{t=m+1}^{m+k} \widehat{\varepsilon}_t = \sum_{t=m+1}^{m+k} (X_t - f(\mathbb{X}_{t-1}, \widehat{\theta}_m)),$$

where $\widehat{\theta}_m$ is the nonlinear least squares estimator for $\widetilde{\theta}_0$ based only on the historic data set $X_1, \ldots, X_m$

$$\widehat{\theta}_m = \arg\min_{\theta\in\Theta} \sum_{t=p+1}^{m} (X_t - f(\mathbb{X}_{t-1}, \theta)\ )^2 \tag{1.6}$$

for a suitable compact set Θ.

The first time it holds

$$|\Gamma(m,k)| \geqslant c\widehat{\sigma}_m g(m,k), \tag{1.7}$$

we stop the monitoring and reject the null hypothesis. Otherwise we continue monitoring.

Here, c is a critical value and $g(m,k)$ is a suitable boundary function, $\widehat{\sigma}_m^2$ is a consistent estimator of $\sigma^2 = \operatorname{var}\varepsilon_1$ in the correctly specified model and for the long-run variance of $\eta_t = Z_t - f(\mathbb{Z}_{t-1}, \widetilde{\theta}_0)$ in the misspecified case. The boundary function is needed in order to obtain a well defined asymptotic distribution of $\max_k \frac{|\Gamma(m,k)|}{g(m,k)}$, where the max is possibly taken over all integers.

In the correctly specified model, a typical estimator of σ^2 is given by

$$\widehat{\sigma}_m^2 = \frac{1}{m - (H(p+2)+1)} \sum_{j=m+1}^{n} (X_t - f(\mathbb{X}_{t-1}, \widehat{\theta}_m))^2, \tag{1.8}$$

which is a consistent estimator in our setup (cf. Lemma 2.1). In simulations this still yields reasonable results for $Z_t = g(\mathbb{Z}_{t-1}) + \varepsilon_t$ with a regression function g that is not a neural network. This is not surprising if the approximation is good enough so that the modeling errors $\eta_t = Z_t - f(\mathbb{Z}_{t-1}, \widetilde{\theta}_0)$ are at least approximately i.i.d. Otherwise an estimator for the long-run variance of η_t is needed.

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.

Consider the stopping time

$$\tau(m) = \begin{cases} \inf\{1 \leqslant k < N(m) :\ \Gamma(m,k) \geqslant c\,\widehat{\sigma}_m\, g(m,k)\}, \\ \infty, \quad \text{if } \Gamma(m,k) < c\,\widehat{\sigma}_m\, g(m,k),\ \text{for all } 1 \leqslant k < N(m), \end{cases}$$

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

In case of the open-end procedure we use a boundary function in the following class

$$g(m,k) = h(m,k,\gamma) = m^{1/2}\left(1+\frac{k}{m}\right)\left(\frac{k}{m+k}\right)^{\gamma}, \qquad 0 \leqslant \gamma < \frac{1}{2}. \tag{1.9}$$

The parameter γ is a tuning parameter, where a γ close to $1/2$ rather detects changes early after the monitoring starts. The motivation behind this specific boundary function is that it leads to a nice asymptotic distribution of $\sup_{k\geqslant 1}\frac{|\Gamma(m,k)|}{g(m,k)}$.

In case of the closed-end procedure the boundary function is in the following class

$$g(m,k) = \sqrt{m}\rho\left(\frac{k}{m}\right), \tag{1.10}$$

where

$$\rho(t) > 0 \qquad \text{continuous on } (0,N],$$
$$\text{such that there exists } 0 \leqslant \gamma < \frac{1}{2}, \quad \lim_{t\to 0} t^{-\gamma}\rho(t) > 0.$$

In particular the conditions are fulfilled for all continuous and positive ρ on $[0,N]$ as well as for $g(m,k) = h(m,k,\gamma)$, $k \leqslant Nm$, with h as in the open-end procedure. As in classical statistics our aim is to control the α-error, i.e.

$$\lim_{m\to\infty} P_{H_0}\left(\tau(m) < \infty\right) = \alpha. \tag{1.11}$$

Unlike in classical statistics the sample size is random and possibly infinite so that it is not suitable for asymptotics. In this setting, the solution proposed by Chu et al. [3] 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 ($\widehat{\theta}_m \xrightarrow{P} \widetilde{\theta}_0$).

Theorem 2.1 shows how to choose the critical value c such that (1.11) holds, i.e. such that the procedure has asymptotic size α. Theorem 2.2 proves that this monitoring procedure detects a large class of alternatives with probability 1 asymptotically, i.e.

$$\lim_{m\to\infty} P_{H_1}\left(\tau(m) < \infty\right) = 1. \tag{1.12}$$

The paper is organized as follows: In Section 2 the null asymptotics of the above sequential tests are given in Theorem 2.1 and the power behavior is obtained in Theorem 2.2. The proofs can be found in Section 3.

2 Consistency of the Change-Point Tests

We will now derive the asymptotics of the monitoring procedures under certain assumptions.

N. 1. Let $\{Z_t : t \in \mathbb{Z}\}$ be a stationary and ergodic process with $\mathbb{E}|Z_1|^\nu < \infty$, for some $\nu \geqslant 2$.

N. 2. Assume that the parameter set $\Theta \subseteq \mathbb{R}^{H(p+2)+1}$ is compact.

N. 3. Let the unique minimizer $\widetilde{\theta}_0$ of $\mathbb{E}(Z_1 - f(\mathbb{Z}_0, \theta))^2$ be an interior point of of Θ.

In the correctly specified model (1.3) the true parameter θ_0 minimizes $\mathbb{E}(Z_1 - f(\mathbb{Z}_0, \theta))^2$ if $\mathbb{E}\varepsilon_1^2 < \infty$. To obtain uniqueness the parameter space has to be chosen in such a way that the network is identifiable. More details on identifiability conditions of neural networks can be found in Kirch and Tadjuidje-Kamgaing [11] as well as Hwang and Ding [9]. In practice this does not play an important role.

N. 4. For $\widehat{\theta}_m$ as in (1.6) and $\widetilde{\theta}_0$ as in (1.4) it holds

$$\sqrt{m}\left(\widehat{\theta}_m - \widetilde{\theta}_0\right) = O_P(1).$$

Kirch and Tadjuidje-Kamgaing [11], Theorem 2.2, give conditions on $\{Z(t)\}$ under which N.4 is fulfilled including a large class of mixing time series.

N. 5. The modeling errors $\eta_t = Z_t - f(\mathbb{Z}_{t-1}, \widetilde{\theta}_0)$ fulfill a strong invariance principle, i.e. there exists a Wiener process $\{W(t)\}$ (possibly after enlarging the probability space) and $\zeta > 0$, $\kappa > 0$ such that

$$\sum_{i=1}^{k}(\eta_i - \mathbb{E}\eta_1) - \kappa W(k) = O\left(k^{-1/2-\zeta}\right) \qquad a.s.$$

In case of a correctly specified model (1.3) $\eta_i = \varepsilon_i$, hence the condition is fulfilled by the classical strong invariance principle with $\kappa^2 = \operatorname{var}\varepsilon_1$ for i.i.d. random variables if $\mathbb{E}|\varepsilon_1|^\nu < \infty$ for some $\nu > 2$ (cf. Komlós et al. and Major [12, 13, 15]).
In the misspecified but well-approximated case the modeling errors $\{\eta_i\}$ can be expected to exhibit only a weak dependency, so that it is reasonable that the assertion still holds. In fact, for strong mixing sequences with $\mathbb{E}|\eta_1|^\nu < \infty$ for some $\nu > 2$ and mixing rate $O(n^{-c})$ for some $c > \nu/(\nu - 2)$, Kuelbs and Philipp [14], Theorem 4, prove the above result. More recently, invariance principles for different dependence concepts have been obtained (cf. e.g. Wu [19]).
In fact, the assertions of Theorems 2.1 and 2.2 can also be derived under the weaker assumption that $\{\eta_i\}$ fulfills a central limit theorem and for $\gamma > 0$ additionally some Hájék-Renyi-type inequality holds. However, proofs become much more technical and involved.

We are now ready to turn to the main results, namely the derivation of the asymptotic null distribution as well as asymptotic power of the sequential tests. The next theorem gives the null asymptotics for the sequential test statistic. It enables us to choose an asymptotic critical value c in such a way that under H_0

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

Theorem 2.2 shows that the corresponding tests have asymptotic power one for a large class of alternatives.

Theorem 2.1. *Assume that N.1 – N.5 and H_0 holds, κ is as in N.5. Then,*

a) for the open-end procedure,

$$\lim_{m\to\infty} P\left(\frac{1}{\kappa}\sup_{1\leqslant k<\infty}\frac{\left|\sum_{t=m+1}^{m+k}\widehat{\varepsilon}_t\right|}{h(m,k,\gamma)} \leqslant c\right) = P\left(\sup_{0<t<1}\frac{|W(t)|}{t^\gamma}\leqslant c\right),$$

where $\{W(t) : t \geqslant 0\}$ denotes a Wiener process,

b) for the closed-end procedure,

$$\lim_{m\to\infty} P\left(\frac{1}{\kappa}\sup_{1\leqslant k<\infty}\frac{\left|\sum_{t=m+1}^{m+k}\widehat{\varepsilon}_t\right|}{\sqrt{m}\rho\left(\frac{k}{m}\right)} \leqslant c\right) = P\left(\sup_{0<t\leqslant N}\frac{|W_1(t)-tW_2(1)|}{\rho(t)}\leqslant c\right),$$

where $\{W_1(t) : t \geqslant 0\}$ and $\{W_2(t) : t \geqslant 0\}$ are independent Wiener processes.

Note that if $\sqrt{m}\rho(k/m) = h(m,k,\gamma)$ the limit in b) converges to the limit in a) for $N \to \infty$ (cf. Horváth et al. [6]). This is the main reason for this choice of boundary function for the open-end procedure. Simulations suggest that the two distributions become very close for $N \geqslant 10$ (cf. Horváth et al. [6], Kirch [10]).

From the above theorem it becomes clear that a consistent estimator for κ is needed to obtain an asymptotic size α test. In the correctly specified case (1.3), it holds $\eta_i = \varepsilon_i$, $\kappa^2 = \operatorname{var}\varepsilon_1$. In this case, Lemma 2.1 shows consistency of the variance estimator (1.8). Otherwise κ is equal to the long-run variance of η_i and is more difficult to estimate. For a discussion of estimators for the long-run variance in the context of change-point analysis we refer to Hušková and Kirch [7].

If the time series is misspecified but well-approximated by the neural network model, then the η_i will be almost i.i.d. and the error from using the estimator $\widehat{\sigma}_m$ may be smaller than the error due to the small sample size.

Lemma 2.1. *Let N.1 – N.5 hold with $2 < \nu < 4$ in N.1. Then, for the correctly specified model* (1.3) *and under H_0 it holds*

$$\widehat{\sigma}_m^2 = \frac{1}{m-(H(p+2)+1)}\sum_{j=p+1}^{m}\widehat{\varepsilon}_j^2 = \sigma^2 + o_p\left(m^{(\nu-2)/\nu}\right).$$

If $\nu \geqslant 4$, then we get the stronger rate $O_P(m^{-1/2})$.

Proof. The proof is analogous to the proof of Lemma 3.1 in Kirch and Tadjuidje-Kamgaing [11]. ■

Under the following alternatives the sequential tests have asymptotic power one, i.e.

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

A. 1. Assume that $\{Y_j : j \geqslant 1\}$ is stationary and ergodic with $\mathbb{E}|Y_1| < \infty$.

A. 2. It holds

$$\left|\mathbb{E}Y_0 - \mathbb{E}f(\mathbb{Y}_p, \widetilde{\theta}_0)\right| > 0.$$

The conditions essentially says that the expectation of the process after the change is different from the expectation of $f(\mathbb{Y}_p, \widetilde{\theta}_0)$, which is the neural network best approximating the time series $\{Z_t\}$ before the change. It becomes clear that even in the correctly specified case detectability goes along with mean changes. This is also confirmed by the simulations although a change from a process to a different process with the same mean are frequently at least unbiased.This is typical in situations, where the statistic is based on sums of estimated residuals and is even true in a simple linear regression situation (cf. Hušková and Koubkova [8]). A larger class of alternatives can usually be detected by using vector-weighted sums of residuals. However, the theory as well as computations necessary become much more involved.

In case of the trivial neural network, where $H = 0$, the condition reduces to $\mathbb{E}Y_1 - \mathbb{E}Z_1 \neq 0$. In this special case, our procedure reduces to $\widehat{\theta}_m = f(x, \theta_m) = \bar{X}_m$, $\widetilde{\theta}_0 = \mathbb{E}Z_1$ and will detect a change to $\{Y_t\}$ if $\mathbb{E}Y_1 \neq \mathbb{E}Z_1$ as long as $\{Y(\cdot)\}$ and $\{Z(\cdot)\}$ are stationary and ergodic sequences as stated above.

Theorem 2.2. *Assume that N.1 – N.5 holds for the autoregressive process before the change-point as well as A.1 and A.2 for the process after the change-point. For the closed-end procedure additionally assume that $k^* = \lfloor \lambda N m \rfloor$, for some $0 \leqslant \lambda < 1$, i.e. the change-point happens before we stop monitoring.*

Then, under the alternative, it holds for all $c > 0$

$$\lim_{m\to\infty} P\left(\sup_{1\leqslant k<N(m)} \frac{\left|\sum_{t=m+1}^{m+k} \widehat{\varepsilon}_t\right|}{g(m,k)} > c\right) = 1.$$

for the open-end procedure as well as closed-end procedure.

3 Proofs

The following lemma is needed in order to prove the main theorems.

Lemma 3.1. *a) Assume that N.1 – N.4 and H_0 hold. Then, as $m \to \infty$,*

$$\sup_{1 \leqslant k < N(m)} \frac{\left| \sum_{i=m+1}^{m+k} \widehat{\varepsilon}_i - \left(\sum_{j=m+1}^{m+k} \eta_j - \frac{k}{m-p} \sum_{j=p+1}^{m} \eta_j \right) \right|}{g(m,k)} = o_P(1),$$

where $\eta_t = Z_t - f(\mathbb{Z}_{t-1}, \widetilde{\theta}_0)$.

b) If additionally N.5 holds, we get

$$\sup_{1 \leqslant k < N(m)} \frac{\left| \left(\sum_{j=m+1}^{m+k} \eta_j - \frac{k}{m-p} \sum_{j=p+1}^{m} \eta_j \right) - (W_{1,m}(k) - \frac{k}{m-p} W_{2,m}(m-p)) \right|}{g(m,k)} = o_P(1),$$

where $\{W_{1,m}(\cdot)\}$ and $\{W_{2,m}(\cdot)\}$ are independent Wiener processes.

Proof. We start with the proof of a). By N.3 and N.4 $\widehat{\theta}_m$ is eventually in the interior of Θ. Hence with probability converging to one it holds

$$\frac{\partial}{\partial \theta} \left(\sum_{t=p+1}^{m} (X_t - f(\mathbb{X}_{t-1}, \theta))^2 \right) \Big|_{\theta = \widehat{\theta}_m} = 0,$$

which implies for m large enough

$$\sum_{t=p+1}^{m} \widehat{\varepsilon}_t = 0. \tag{3.1}$$

From this we can conclude

$$\begin{aligned}
&\sum_{t=m+1}^{m+k} \widehat{\varepsilon}_t - \left(\sum_{i=m+1}^{m+k} \eta_i - \frac{k}{m-p} \sum_{i=p+1}^{m} \eta_i \right) = \sum_{t=m+1}^{m+k} (\widehat{\varepsilon}_t - \eta_t) - \frac{k}{m-p} \sum_{t=p+1}^{m} (\widehat{\varepsilon}_t - \eta_t) \\
&= \sum_{t=m+1}^{m+k} (f(\mathbb{Z}_{t-1}, \widetilde{\theta}_0) - f(\mathbb{Z}_{t-1}, \widehat{\theta}_m)) - k \mathbb{E} \nabla f(\mathbb{Z}_p, \widetilde{\theta}_0)^T (\widetilde{\theta}_0 - \widehat{\theta}_m) \\
&\qquad - \frac{k}{m-p} \sum_{t=p+1}^{m} (f(\mathbb{Z}_{t-1}, \widetilde{\theta}_0) - f(\mathbb{Z}_{t-1}, \widehat{\theta}_m)) + k \mathbb{E} \nabla f(\mathbb{Z}_p, \widetilde{\theta}_0)^T (\widetilde{\theta}_0 - \widehat{\theta}_m) \\
&=: D_1(m,k) - D_2(m,k).
\end{aligned}$$

A Taylor expansion of f yields

$$\begin{aligned}
&f(\mathbb{Z}_{t-1}, \widehat{\theta}_m) - f(\mathbb{Z}_{t-1}, \widetilde{\theta}_0) \\
&= \nabla f(\mathbb{Z}_{t-1}, \widetilde{\theta}_0)^T (\widehat{\theta}_m - \widetilde{\theta}_0) + \frac{1}{2} (\widehat{\theta}_m - \widetilde{\theta}_0)^T \nabla^2 f(\mathbb{Z}_{t-1}, \xi) (\widehat{\theta}_m - \widetilde{\theta}_0),
\end{aligned} \tag{3.2}$$

where $\nabla f(\mathbb{Z}_{t-1}, \theta)$ is the gradient with respect to θ and $\nabla^2 f(\mathbb{Z}_{t-1}, \theta)$ is the Hessian matrix, ξ is between $\widetilde{\theta}_0$ and $\widehat{\theta}_m$ elementwise. Furthermore the Hessian matrix is by Assumption N.2 and the twice continuous differentiability of f uniformly bounded by

$O(1)\ \max_{1\leqslant i\leqslant p}\max_{1\leqslant j\leqslant p}|Z_{t-i}Z_{t-j}|$, the gradient by $O(1)\ \max_{1\leqslant j\leqslant p}|Z_{t-j}|$. By a uniform law of large numbers for stationary and ergodic processes (cf. Ranga Rao [16], Theorem 6.5)

$$\sup_{\xi\in K}\sup_{p<k\leqslant n}\frac{1}{k}\sum_{t=p+1}^{k}\|\nabla^2 f(\mathbb{Z}_{t-1},\xi)\|_\infty = O_P(1),$$

where $\|(\alpha_{i,j})\|_\infty = \max_{i,j}|\alpha_{i,j}|$. Together with (3.2) this yields uniformly in k

$$\begin{aligned}&\sum_{t=p+1}^{k}\left(f(\mathbb{Z}_{t-1},\widehat{\theta}_m)-f(\mathbb{Z}_{t-1},\widetilde{\theta}_0)\right)\\ &=\sum_{t=p+1}^{k}\nabla f(\mathbb{Z}_{t-1},\widetilde{\theta}_0)^T(\widehat{\theta}_m-\widetilde{\theta}_0)+O_P\left(k\|\widehat{\theta}_m-\widetilde{\theta}_0\|^2\right).\end{aligned}\tag{3.3}$$

An application of the ergodic theorem yields

$$\frac{1}{l}\sum_{t=p+1}^{l}\left(\nabla f(\mathbb{Z}_{t-1},\widetilde{\theta}_0)^T-\mathbb{E}\nabla f(\mathbb{Z}_{t-1},\widetilde{\theta}_0)^T\right)=o(1)\quad a.s.\qquad (l\to\infty).\tag{3.4}$$

Note that for some $C>0$

$$h(m,k,\gamma)\geqslant\begin{cases}Cm^{1/2-\gamma}\,k^{\gamma}, & k\leqslant m,\\ Cm^{-1/2}\,k, & k>m.\end{cases}\tag{3.5}$$

This together with (3.3), (3.4) and N.4 yields (as $m\to\infty$)

$$\sup_{k\geqslant 1}\frac{|D_1(m,k)|}{h(m,k,\gamma)}=O_P(1)\sup_{k\leqslant\sqrt{m}}\left(\frac{k}{m}\right)^{1-\gamma}+o_P(1)\sup_{\sqrt{m}<k\leqslant m}\left(\frac{k}{m}\right)^{1-\gamma}+o_P(1)=o_P(1).\tag{3.6}$$

Similarly

$$\sup_{k\geqslant 1}\frac{|D_2(m,k)|}{h(m,k,\gamma)}=o_P(1).\tag{3.7}$$

The assertion in a) follows from (3.6) and (3.7) for the open-end procedure, similar arguments lead to the assertion for the closed-end procedure.

Concerning b) let $1>\xi>\max(0,1-\zeta/\gamma)$ for $\gamma\neq 0$ and $0<\xi<1$ for $\gamma=0$. Furthermore $\widetilde{\eta}_i=\eta_i-\mathbb{E}\eta_1$. By (3.5) and N.5 we obtain

$$\begin{aligned}&\sup_{k>m}\frac{\left|\sum_{i=m+1}^{m+k}\widetilde{\eta}_i-\kappa\left(W(m+k)-W(m)\right)\right|}{h(m,k,\gamma)}=O_P(1)\sup_{k>m}\frac{m^{1/2}(m+k)^{1/2-\zeta}}{k}\\ &=O_P\left(m^{-\zeta}\right)=o_P(1).\end{aligned}\tag{3.8}$$

Similarly

$$\sup_{m^\xi \leqslant k \leqslant m} \frac{\left|\sum_{i=m+1}^{m+k} \widetilde{\eta}_i - \kappa\left(W(m+k) - W(m)\right)\right|}{h(m,k,\gamma)} = O_P(1) \sup_{m^\xi \leqslant k \leqslant m} \frac{m^\gamma (m+k)^{1/2-\zeta}}{m^{1/2} k^\gamma}$$
$$= O_P\left(m^{-(\zeta-\gamma+\xi\gamma)}\right) = o_P(1), \tag{3.9}$$

as by definition of ξ it holds $\zeta - \gamma + \xi\gamma > 0$. For $k < m^\xi$ first note that by the law of iterated logarithm

$$W(m+k) - W(m) \overset{\mathcal{D}}{=} W(k) = O_P(\sqrt{k \log\log k})$$

as well as

$$\sum_{i=m+1}^{m+k} \widetilde{\eta}_i \overset{\mathcal{D}}{=} \sum_{i=1}^{k} \widetilde{\eta}_i = O_P(k^{1/2-\zeta}) + \kappa W(k) = O_P(\sqrt{k \log\log k}).$$

From this we can conclude by (3.5) as $\xi < 1, \gamma < 1/2$

$$\sup_{1 \leqslant k < m^\xi} \frac{\left|\sum_{i=m+1}^{m+k} \widetilde{\eta}_i - \kappa\left(W(m+k) - W(m)\right)\right|}{h(m,k,\gamma)} = O_P(1) \sup_{1 \leqslant k < m^\xi} \frac{(k \log\log k)^{1/2} m^\gamma}{m^{1/2} k^\gamma}$$
$$= O_P(1) m^{(\xi-1)(1/2-\gamma)} \log\log m = o_P(1). \tag{3.10}$$

Similar arguments yield

$$\sup_{k \geqslant 1} \frac{k}{m\, h(m,k,\gamma)} \left| \sum_{i=p+1}^{m} \widetilde{\eta}_i - \kappa W(m) \right| = O_P(m^{-\zeta}) = o_P(1). \tag{3.11}$$

Putting together (3.8) – (3.11) yields the assertion for the open-end procedure noting that $W_{1,m}(t) = W(m+t) - W(m), t \geqslant 0$, and $W_{2,m}(t), 0 \leqslant t \leqslant m$, are independent Wiener processes. The arguments for the closed-end procedure are similar. ■

We can now easily deduce the assertion of Theorem 2.1.

Proof of Theorem 2.1. The assertion follows from Lemma 3.1 in the same way as Theorem 2.1 in Horváth et al. [6] is derived from their Lemma 5.3. There, the assertion is proven for the open-end procedure only. However, the proof remains true for closed-end procedures with a boundary function as given here. ■

Finally we prove Theorem 2.2.

Proof of Theorem 2.2. For $\widetilde{k} > k^*$ it holds

$$\sum_{j=p+1}^{m+\widetilde{k}} \widehat{\varepsilon}_j$$
$$= \sum_{j=1}^{m+k^*} \left(Z_j - f(\mathbb{Z}_{j-1}, \widehat{\theta}_m)\right) + \sum_{j=m+k^*+1}^{m+k^*+p} (Y_j - f(\mathbb{X}_{j-1}, \widehat{\theta}_m))$$
$$+ \sum_{j=m+k^*+p+1}^{m+\widetilde{k}} \left(f(\mathbb{Y}_{j-1}, \widetilde{\theta}_0) - f(\mathbb{Y}_{j-1}, \widehat{\theta}_m)\right) + \sum_{j=m+k^*+p+1}^{m+\widetilde{k}} \left(Y_j - f(\mathbb{Y}_{j-1}, \widetilde{\theta}_0)\right)$$
$$=: D_3(m,k^*) + D_4(m,k^*) + D_5(m,k*,\widetilde{k}) + D_6(m,k^*,\widetilde{k}). \tag{3.12}$$

By Theorem 2.1 it holds

$$\max_{1\leqslant k<N(m)} \frac{D_3(m,k)}{g(m,k)} = O_P(1). \tag{3.13}$$

For the open-end procedure let $\widetilde{k} = k^* + \max(k^*, m)$. By the compactness assumption of Θ and the boundedness of the network function ψ it holds

$$\sup_{\theta\in\Theta} \sup_{\mathbf{x}\in\mathbb{R}^p} |f(\mathbf{x},\theta)| \leqslant C < \infty. \tag{3.14}$$

Hence $\sum_{j=m+k^*+1}^{m+k^*+p} f(\mathbb{X}_{j-1}, \widehat{\theta}_m) \leqslant Cp$ and by the stationarity of $\{Y(\cdot)\}$ it holds $\sum_{j=m+k^*+1}^{m+k^*+p} Y_j \overset{\mathcal{D}}{=} \sum_{j=1}^{p} Y_j = O_P(1)$, hence by (3.5)

$$\frac{D_4(m,k^*,\widetilde{k})}{h(m,\widetilde{k},\gamma)} = O_P(1)\,\frac{1}{h(m,\widetilde{k},\gamma)} = O_P\left(\frac{1}{\sqrt{m}}\right). \tag{3.15}$$

By (3.3), (3.4) and N.4

$$\frac{D_5(m,k^*,\widetilde{k})}{h(m,\widetilde{k},\gamma)} = O_P\left(\frac{\widetilde{k}-k^*}{\widetilde{k}}\right) = O_P(1). \tag{3.16}$$

$\{Y_t - f(\mathbb{Y}_t, \widetilde{\theta}_0); t\}$ is stationary and ergodic with $\mathbb{E}|Y_t - f(\mathbb{Y}_t, \widetilde{\theta}_0)| < \infty$ by (3.14). Hence, the ergodic theorem yields

$$D_6(m,k^*,\widetilde{k}) = (\widetilde{k}-k^*)(\mathbb{E}Y_0 - \mathbb{E}f(\mathbb{Y}_p,\widetilde{\theta}_0)) + o_P\left(\widetilde{k}-k^*\right) \tag{3.17}$$

and by (3.5)

$$\frac{\widetilde{k}-k^*}{h(m,\widetilde{k},\gamma)} = O(\sqrt{m}).$$

Since

$$h(m,\widetilde{k},\gamma) \leqslant 2\frac{\widetilde{k}}{\sqrt{m}},$$

we can conclude

$$\begin{aligned}\frac{D_6(m,\widetilde{k})}{h(m,\widetilde{k},\gamma)} &\geqslant \frac{(\widetilde{k}-k^*)\sqrt{m}}{2\widetilde{k}}(\mathbb{E}Y_0 - \mathbb{E}f(\mathbb{Y}_p,\widetilde{\theta}_0)) + o_P\left(\sqrt{m}\right)\\ &\geqslant \frac{\sqrt{m}}{4}(\mathbb{E}Y_0 - \mathbb{E}f(\mathbb{Y}_p,\widetilde{\theta}_0)) + o_P\left(\sqrt{m}\right).\end{aligned} \tag{3.18}$$

Putting together (3.12) –(3.18) and Assumption A.2 yields

$$\sup_{1\leqslant k<\infty} \frac{\left|\sum_{i=m+1}^{m+k}\widehat{\varepsilon}_i\right|}{h(m,k,\gamma)} \geqslant \frac{\sqrt{m}}{4}(\mathbb{E}Y_0 - \mathbb{E}f(\mathbb{Y}_p,\widetilde{\theta}_0)) + o_P\left(\sqrt{m}\right) \overset{P}{\longrightarrow} \infty, \tag{3.19}$$

which concludes the proof for the open-end procedure. Similar arguments using $\widetilde{k} = Nm$ yield the assertion for the closed-end procedure. ■

References

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

[2] Aue, A., Hörmann, S., Horváth, L., and Hušková, M. Sequential testing for the stability of portfolio betas. 2009. In preparation.

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

[4] Franke, J. and Mabouba, D. Estimating market risk with neural networks. *Statist. Decisions*, 30:63–82, 2006.

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

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

[7] Hušková, M., and Kirch, C. A note on studentized confidence intervals in change-point analysis. *Comput. Statist.*, 25:269–289, 2010.

[8] Hušková, M. and Koubková, A. Monitoring jump changes in linear models. *J. Statist. Res.*, 39:59–78, 2005.

[9] Hwang, J. T. G. and Ding, A. A. Prediction intervals for artificial neural networks. *J. Amer. Stat. Assoc.*, 92:748–757, 1997.

[10] Kirch, C. Bootstrapping sequential change-point tests. *Sequential Anal.*, 27:330–349, 2008.

[11] Kirch, C. and Tadjuidje K., J. . Testing for parameter stability in nonlinear autoregressive models. *Preprint*, 2011.

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

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

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

[15] Major, P. An approximation of partial sums of independent rvs. *Z. Wahrsch. verw. Geb.*, 35:213–220, 1976.

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

[17] Stockis, J.-P., Franke, J., and Tadjuidje K., J. On geometric ergodicity of CHARME models. *J. Time Ser. Anal.*, 31:141–152, 2010.

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

[19] Wu, W. B. Strong invariance principles for dependent random variables. *Ann. Probab.*, 35:2294–2320, 2007.